02 109
RSW

日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年　７月１９日

出願番号
Application Number: 特願２００２－２１１６３４

[ST.10/C]: [ＪＰ２００２－２１１６３４]

出願人
Applicant(s): インターナショナル・ビジネス・マシーンズ・コーポレーション

２００２年１２月１３日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００２－３０９８２７５

【書類名】　特許願

【整理番号】　JP9020109

【提出日】　平成14年 7月19日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 17/30

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　野美山　浩

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　岩男　俊孝

【特許出願人】

【識別番号】　390009531

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】　100086243

【弁理士】

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【弁理士】

【氏名又は名称】　市位　嘉宏

【代理人】

【識別番号】　100108501

【弁理士】

【氏名又は名称】　上野　剛史

【復代理人】

【識別番号】　100104880

【弁理士】

【氏名又は名称】　古部　次郎

【手数料の表示】

【予納台帳番号】　081504

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9706050

【包括委任状番号】　9704733

【包括委任状番号】　0004480

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　情報検索システム、情報検索方法、ＨＴＭＬ文書の構造解析方法及びプログラム

【特許請求の範囲】

【請求項１】　ネットワークを介してウェブサイトのクローリングを行う情報検索システムにおいて、

所定のウェブページにおける意味を考慮してソースコードの構造を解析する構造解析部と、

前記構造解析部の解析結果に基づいて、前記所定のウェブページからリンクされる他のウェブサイトの重要度を計算する重要度計算部と、

前記重要度計算部により計算された重要度に応じてウェブサイトをクロールするクローリング実行部と

を備えることを特徴とする情報検索システム。

【請求項２】　前記構造解析部は、前記ソースコードに含まれる情報要素のうち、相互に関連する情報要素を対応付けることを特徴とする請求項１に記載の情報検索システム。

【請求項３】　前記重要度計算部は、前記ウェブサイトの重要度を計算するための戦略を、予め用意された戦略の中から選択的に用いて重要度の計算を行うことを特徴とする請求項１に記載の情報検索システム。

【請求項４】　前記重要度計算部は、前記ウェブサイトの重要度を計算するための戦略として、複数の戦略を選択し、各々重みを付けて用いることを特徴とする請求項３に記載の情報検索システム。

【請求項５】　ＨＴＭＬ文書の文書構造を意味を考慮して解析し、解析によって得られた情報要素を対応するアンカーに付加する文書構造解析部と、

前記文書構造解析部の解析により得られた前記情報要素に基づいて計算された前記アンカーの重要度に応じて当該アンカーにてリンクされるウェブサイトをクロールするクローリング実行部と

を備えることを特徴とする情報検索システム。

【請求項６】　前記文書構造解析部は、前記ＨＴＭＬ文書を構成する各情報

要素を、当該情報要素が持つ意味のまとまりごとにブロック化し、各ブロック内の情報要素を付加情報として同一ブロック内のアンカーに付加することを特徴とする請求項５に記載の情報検索システム。

【請求項７】　前記文書構造解析部の解析により得られた前記情報要素に基づき、予め選択された所定の戦略にしたがって前記アンカーの重要度を計算する重要度計算部をさらに備え、

前記クローリング実行部は、前記重要度計算部にて計算された前記アンカーの重要度に応じてウェブサイトをクロールすることを特徴とする請求項５に記載の情報検索システム。

【請求項８】　コンピュータを用いて、ネットワークを介してウェブサイトのクローリングを行う情報検索方法であって、

初期情報となるウェブページを取得してソースコードを記憶装置に格納するステップと、

前記記憶装置から前記ウェブページのソースコードを読み出し、当該ウェブページにおける意味を考慮して構造解析を行い、解析結果を前記記憶装置に格納するステップと、

前記記憶装置に格納された前記構造解析の結果に基づいて、前記ウェブページからリンクされる他のウェブサイトの重要度を計算するステップと、

計算された重要度に応じてウェブサイトにアクセスし、コンテンツを取得して前記記憶装置に格納するステップと

を含むことを特徴とする情報検索方法。

【請求項９】　メモリから処理対象であるＨＴＭＬ文書を読み出し、当該ＨＴＭＬ文書を構成する各情報要素を、当該ＨＴＭＬ文書のタグに基づいてブロック化し、ブロック化された当該ＨＴＭＬ文書の構造データをメモリに格納する第１のステップと、

前記メモリからブロック化された前記ＨＴＭＬ文書の構造データを読み出し、意味的に相互に関連する情報要素を対応付けることにより、当該ＨＴＭＬ文書のブロック構造を更新し、更新された当該構造データをメモリに格納する第２のステップと

を含むことを特徴とするコンピュータを用いたＨＴＭＬ文書の構造解析方法。

【請求項１０】　ネットワークに接続されたコンピュータを制御して、ウェブサイトのクローリングを行うプログラムであって、

初期情報となるウェブページを取得してソースコードを記憶装置に格納する処理と、

前記記憶装置から前記ウェブページのソースコードを読み出し、当該ウェブページにおける意味を考慮して構造解析を行い、解析結果を前記記憶装置に格納する処理と、

前記記憶装置に格納された前記構造解析の結果に基づいて、前記ウェブページからリンクされる他のウェブサイトの重要度を計算する処理と、

計算された重要度に応じてウェブサイトにアクセスし、コンテンツを取得して前記記憶装置に格納する処理と

を前記コンピュータに実行させることを特徴とするプログラム。

【請求項１１】　前記プログラムは、前記ソースコードに含まれる情報要素のうち、相互に関連する情報要素を対応付けることにより、前記構造解析を前記コンピュータに実行させることを特徴とする請求項１０に記載のプログラム。

【請求項１２】　前記プログラムによる前記ウェブサイトの重要度を計算する処理では、前記ウェブサイトの重要度を計算するための戦略として、複数の戦略を各々重みを付けて用いることを特徴とする請求項１０に記載のプログラム。

【請求項１３】　コンピュータを制御して、ＨＴＭＬ文書の文書構造を解析するプログラムであって、

メモリから処理対象であるＨＴＭＬ文書を読み出し、当該ＨＴＭＬ文書を構成する各情報要素を、当該ＨＴＭＬ文書のタグに基づいてブロック化し、ブロック化された当該ＨＴＭＬ文書の構造データをメモリに格納する第１の処理と、

前記メモリからブロック化された前記ＨＴＭＬ文書の構造データを読み出し、意味的に相互に関連する情報要素を対応付けることにより、当該ＨＴＭＬ文書のブロック構造を更新し、更新された当該構造データをメモリに格納する第２の処理と

を前記コンピュータに実行させることを特徴とするプログラム。

【請求項１４】　前記プログラムによる前記第２の処理では、

文書構造解析の目的に鑑みて不要な情報要素を識別する処理と、

構造的に意味のないブロックを削除する処理と、

前記情報要素の内容に基づいて、情報要素のマージあるいはブロックの分割を行う処理と、

各ブロックに含まれる情報に基づいて、ブロック構造をマージする処理と

を前記コンピュータに実行させることを特徴とする請求項１３に記載のプログラム。

【請求項１５】　ネットワークに接続されたコンピュータを制御してウェブサイトのクローリングを行うプログラムを、当該コンピュータが読み取り可能に記録した記録媒体において、

前記プログラムは、

初期情報となるウェブページを取得してソースコードを記憶装置に格納する処理と、

前記記憶装置から前記ウェブページのソースコードを読み出し、当該ウェブページにおける意味を考慮して構造解析を行い、解析結果を前記記憶装置に格納する処理と、

前記記憶装置に格納された前記構造解析の結果に基づいて、前記ウェブページからリンクされる他のウェブサイトの重要度を計算する処理と、

計算された重要度に応じてウェブサイトにアクセスし、コンテンツを取得して前記記憶装置に格納する処理と

を前記コンピュータに実行させることを特徴とする記録媒体。

【請求項１６】　コンピュータを制御してＨＴＭＬ文書の文書構造を解析するプログラムを、当該コンピュータが読み取り可能に記録した記録媒体において、

前記プログラムは、

メモリから処理対象であるＨＴＭＬ文書を読み出し、当該ＨＴＭＬ文書を構成する各情報要素を、当該ＨＴＭＬ文書のタグに基づいてブロック化し、ブロック化された当該ＨＴＭＬ文書の構造データをメモリに格納する第１の処理と、

前記メモリからブロック化された前記ＨＴＭＬ文書の構造データを読み出し、意味的に相互に関連する情報要素を対応付けることにより、当該ＨＴＭＬ文書のブロック構造を更新し、更新された当該構造データをメモリに格納する第２の処理と

を前記コンピュータに実行させることを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ネットワークを介して所望の情報を自動的に取得する技術に関し、特にインターネット上で提供されているウェブコンテンツを検索し、リンクを辿りながら取得（クローリング）する技術に関する。

【０００２】

【従来の技術】

今日、インターネットに代表されるコンピュータのネットワーク環境が広く普及したことにより、ネットワーク上で提供されている膨大な情報の中から検索エンジンを用いて所望の情報を検索し取得することが一般的に行われている。この検索エンジンには多くの種類が存在するが、予め情報を検索して取得しておき、検索要求に応じて保持している情報を返す静的な検索エンジンを用いた場合は、膨大な情報源（ウェブページ等）を対象としなければならないため、最新の情報を獲得することが困難である。また、基本的に検索エンジンを持つサーバが全ての処理を行うことが前提となっているため、サーバの負担が大きい。

【０００３】

そこで、静的な検索エンジンで集めたキーワード検索結果の集合を初期集合として用い、これを起点に関連するサイトを動的に検索する手法が提案されている。この種の従来技術としては、例えば、下記文献１に開示されたShark-Searchと呼ばれる検索技術がある。

文献１：Michael Herscovici, Michal Jacovi, Yoelle S. Maarek, Dan Pelleg, Menachem Shtalhaim, Sigalit Ur. "The Shark-Search Algorithm: An Applica

tion: Tailored Web SiteMapping" In the Proceedings of WWW7, the 7th International World Wide Web Conference, Brisbane, April 1998. Also appeared in the Journal of Computer Networks and ISDN 30 (1998), pp 317-326. HYPERLINK "http://www7.scu.edu.au/programme/fullpapers/1849/com1849.htm" http://www7.scu.edu.au/programme/fullpapers/1849/com1849.htm

【０００４】

同文献に開示された技術は、インターネット上で、指定されたＵＲＬ（Uniform Resource Locator）とキーワードとに基づき、当該ＵＲＬのウェブサイトから当該キーワードに関連するウェブサイト（重要度の高いウェブサイト）を動的に検索する。このシステムは、初期集合を求めるためのキーワード（domain query）と動的にウェブサイトをクロールする際にクロール対象であるウェブサイトの重要度の計算に用いるキーワード（focused query）の２つを用いることによって精度の向上を図っている。

【０００５】

【発明が解決しようとする課題】

上述したように、ネットワーク上で提供される膨大な最新の情報を効率よく検索するために、検索要求があった場合に動的に情報検索を行うことが求められる。

しかし、上述した従来の動的な検索エンジンは、基本的に、ユーザが指定したトピック（キーワード等）に近い情報という１つの判断基準（relevance()）で検索を行う。そのため、情報の使用目的に応じて多様な戦略で柔軟に検索を行うことができなかった。

【０００６】

また、情報を効率的に検索するためには、取得対象である情報（ウェブページ等）の重要度を判断し、これに基づいて情報の取得順や取得範囲を決める必要がある。しかし、インターネット上でＵＲＬ及びトピックに基づいてウェブサイトをクロールする従来技術では、この重要度を効果的に判断することができなかった。すなわち、情報の重要度を判断するために限られた情報、例えばウェブペー

ジにおける、指定されたキーワードや、アンカーに近い位置に記載されているテキストなどしか用いておらず、所望の情報を効率的に検索することができなかった。例えば、上記文献１に開示された従来技術の場合、文献１には、アンカーの重要性を判断するために当該アンカーの近傍のテキスト（anchor_text_context）を考慮するとの記述があるが、どのようにしてこの anchor_text_contextを得るかについて明確な記述はない。

【０００７】

そこで、本発明は、情報の使用目的に応じて多様な戦略による柔軟な情報検索を可能とすることを目的とする。

また、本発明は、ウェブサイトのクローリングにおいて、この多様な戦略による情報検索を実現するために、ウェブページに含まれる情報を有効に活用して検索を行うことを目的とする。

【０００８】

【課題を解決するための手段】

上記の目的を達成する本発明は、ネットワークを介してウェブサイトのクローリングを行う、次のように構成された情報検索システムとして実現される。すなわち、この情報検索システムは、所定のウェブページにおける意味を考慮してソースコードの構造を解析する構造解析部と、この構造解析部の解析結果に基づいて、このウェブページからリンクされる他のウェブサイトの重要度を計算する重要度計算部と、この重要度計算部により計算された重要度に応じてウェブサイトをクロールするクローリング実行部とを備えることを特徴とする。

より詳しくは、この重要度計算部は、ウェブサイトの重要度を計算するための戦略を、予め用意された戦略の中から選択的に用いて重要度の計算を行う。さらに好ましくは、この重要度計算部は、複数の戦略を選択し、各々重みを付けて用いる。

【０００９】

また、本発明による他の情報検索システムは、ＨＴＭＬ文書の文書構造を意味を考慮して解析し、解析によって得られた情報要素を対応するアンカーに付加する文書構造解析部と、この文書構造解析部の解析により得られた情報要素に基づ

いて計算されたアンカーの重要度に応じて、このアンカーにてリンクされるウェブサイトをクロールするクローリング実行部とを備えることを特徴とする。

ここで詳細には、この文書構造解析部は、ＨＴＭＬ文書を構成する各情報要素を、この情報要素が持つ意味のまとまりごとにブロック化し、各ブロック内の情報要素を付加情報として同一ブロック内のアンカーに付加する。

この情報検索システムは、文書構造解析部の解析により得られた情報要素に基づき、予め選択された所定の戦略にしたがってアンカーの重要度を計算する重要度計算部をさらに備えることができる。これにより、クローリング実行部は、重要度計算部にて所定の戦略にしたがって計算されたアンカーの重要度に応じてウェブサイトをクロールすることができる。

【００１０】

また、上記の目的を達成する本発明は、コンピュータを用いネットワークを介してウェブサイトのクローリングを行う、次のような情報検索方法として実現される。この情報検索方法は、初期情報となるウェブページを取得してソースコードを記憶装置に格納するステップと、この記憶装置からウェブページのソースコードを読み出し、このウェブページにおける意味を考慮して構造解析を行うステップと、この構造解析の結果に基づいて、このウェブページからリンクされる他のウェブサイトの重要度を計算するステップと、計算された重要度に応じてウェブサイトにアクセスし、コンテンツを取得するステップとを含むことを特徴とする。

【００１１】

さらに、本発明は、この情報検索方法などで用いられる、次のようなＨＴＭＬ文書の構造解析方法としても実現される。すなわち、このＨＴＭＬ文書の構造解析方法は、処理対象であるＨＴＭＬ文書を構成する各情報要素を、このＨＴＭＬ文書のタグに基づいてブロック化するステップと、ブロック化された前記ＨＴＭＬ文書の構造データにおいて、意味的に相互に関連する情報要素を対応付けることにより、このＨＴＭＬ文書のブロック構造を更新するステップとを含む。

より詳細には、このＨＴＭＬ文書のブロック構造を更新するステップは、文書構造解析の目的に鑑みて不要な情報要素を識別するステップと、構造的に意味の

ないブロックを削除するステップと、この情報要素の内容に基づいて、情報要素のマージあるいはブロックの分割を行うステップと、各ブロックに含まれる情報に基づいて、ブロック構造をマージするステップとを含む。

【００１２】

また、本発明は、コンピュータを制御して上述した情報検索システムとして機能させるプログラムや、上述した情報検索方法またはＨＴＭＬ文書の構造解析方法における各ステップに対応する処理をコンピュータに実行させるプログラムとして実現することができる。このプログラムは、磁気ディスクや光ディスク、半導体メモリ、その他の記録媒体に格納して配布したり、ネットワークを介して配信したりすることにより、提供することができる。

【００１３】

【発明の実施の形態】

以下、添付図面に示す実施の形態に基づいて、この発明を詳細に説明する。

図１は、本実施の形態による情報検索システムを実現するのに好適なコンピュータ装置のハードウェア構成の例を模式的に示した図である。

図１に示すコンピュータ装置は、演算手段であるＣＰＵ（Central Processing Unit：中央処理装置）１０１と、Ｍ／Ｂ（マザーボード）チップセット１０２及びＣＰＵバスを介してＣＰＵ１０１に接続されたメインメモリ１０３と、同じくＭ／Ｂチップセット１０２及びＡＧＰ（Accelerated Graphics Port）を介してＣＰＵ１０１に接続されたビデオカード１０４と、ＰＣＩ（Peripheral Component Interconnect）バスを介してＭ／Ｂチップセット１０２に接続されたハードディスク１０５、ネットワークインターフェイス１０６及びＵＳＢポート１０７と、さらにこのＰＣＩバスからブリッジ回路１０８及びＩＳＡ（Industry Standard Architecture）バスなどの低速なバスを介してＭ／Ｂチップセット１０２に接続されたフロッピーディスクドライブ１０９及びキーボード／マウス１１０とを備える。

なお、図１は本実施の形態を実現するコンピュータ装置のハードウェア構成を例示するに過ぎず、本実施の形態を適用可能であれば、他の種々の構成を取ることができる。例えば、ビデオカード１０４を設ける代わりに、ビデオメモリのみ

を搭載し、ＣＰＵ１０１にてイメージデータを処理する構成としても良いし、ＡＴＡ（AT Attachment）などのインターフェイスを介してＣＤ－ＲＯＭ（Compact Disc Read Only Memory）やＤＶＤ－ＲＯＭ（Digital Versatile Disc Read Only Memory）のドライブを設けても良い。

【００１４】

本実施の形態では、情報としてインターネット上で提供される各種のウェブコンテンツ（ウェブページやそのオブジェクト）を検索し獲得する場合を例として説明する。したがって、本実施の形態において図１に示すコンピュータ装置は、プログラム制御されたＣＰＵ１０１にて実現される通信制御手段及びネットワークインターフェイス１０６を介して、インターネットに接続し、ウェブサイトにアクセスする。

図２は、図１に示したコンピュータ装置にて実現される本実施の形態による情報検索システムの構成を示す図である。

図２に示すように、本実施の形態による情報検索システムは、インターネット上のウェブサイトからリンクを辿り所望の情報に関連のあるウェブサイトを検索する情報取得手段であるクローラ１０と、クローラ１０にて検索されたウェブサイトに対して所定の条件に基づく選別を行うウェブサイト選別部２０と、ウェブサイト選別部２０による選別にて選ばれたウェブサイトに基づいて、種々の戦略に基づくレポートを作成するレポート作成部３０とを備えている。

【００１５】

上述したクローラ１０、ウェブサイト選別部２０及びレポート作成部３０は、図１に示したメインメモリ１０３に展開されたプログラムにてＣＰＵ１０１を制御することにより実現される仮想的なソフトウェアブロックである。ＣＰＵ１０１を制御してこれらの機能を実現させる当該プログラムは、磁気ディスクや光ディスク、半導体メモリ、その他の記憶媒体に格納して配布したり、ネットワークを介して配信したりすることにより提供される。本実施の形態では、図１に示したネットワークインターフェイス１０６やフロッピーディスクドライブ１０８、図示しないＣＤ－ＲＯＭドライブなどを介して当該プログラムを入力し、ハードディスク１０５に格納する。そして、ハードディスク１０５に格納されたプログ

ラムをメインメモリ１０３に読み込んで展開し、ＣＰＵ１０１にて実行することにより、図２に示した各構成要素の機能を実現する。

【００１６】

図３は、本実施の形態の情報検索システムによる情報検索の概略的な流れを示すフローチャートである。

図３に示すように、本実施の形態の情報検索システムは、図２に示したクローラ１０により、初期サイトを獲得し（ステップ３０１）、ユーザによって任意に選択された種々の戦略にしたがって動的にウェブサイトのクローリングを行う（ステップ３０２）。

ここで、初期サイトとは、ウェブサイトのクローリングを開始するために初期的に設定されるウェブサイトもしくはその集合である。これをインデックスとして用いて次のクローリングを行う。また、戦略とは、ウェブサイトのクローリングを行う上での基準となる方針を意味し、具体的には検索条件等として設定される。本実施の形態で採られる戦略については３８段落以下で詳述する。

次に、情報検索システムは、ウェブサイト選別部２０により、クローリングで検索されたウェブサイトの集合の中から、検索条件であるトピックとの関連性や時間的条件に基づいて、有効なウェブサイトを選択する（ステップ３０３）。そして最後に、ウェブサイト選別部２０にて選択されたウェブサイトに対し、上記の戦略に基づいて評価を行い、レポートを作成する（ステップ３０４）。作成されたレポートはウェブページ等の形でディスプレイ装置に表示され、あるいはハードディスク１０５等の記憶装置に保存される。

【００１７】

本実施の形態において、クローラ１０は、図２に示したように、初期サイトを取得する初期サイト獲得部１１と、当該初期サイトに対応するウェブページに対して文書構造解析を行う文書構造解析部１２と、文書構造解析部１２による解析結果に基づいてクローリングによる取得対象であるウェブサイトの重要度を計算する重要度計算部１３と、クローリングによるウェブサイトの取得処理を実行するクローリング実行部１４とを備える。

初期サイト獲得部１１における初期サイトの獲得には、例えば次のような方法

を採ることができる。

・情報を収集しようとする特定のウェブサイト（例えば企業サイト）のホームページ（トップページ）のＵＲＬを指定。

・任意のキーワードに対して既存の検索エンジンを用いて検索。

既存の検索エンジンとしては、インターネット上で提供される一般的な検索サービスを利用することができる。検索エンジンを用いて初期サイトを取得した場合は、検索によって取得されたウェブサイトの集合が初期サイト（初期サイト集合）となる。

【００１８】

ウェブサイトのクローリングは、初期サイトを起点として、そこからアンカータグなどの情報に基づいて参照されているウェブサイトを獲得し、その中から指定された戦略に一致するウェブサイトを獲得する処理である。この処理は、獲得されたウェブサイトの数の上限、深さの上限、クロールの時間制限などの終了条件を予め設定しておき、この終了条件を満たすまで再帰的に適用される。ユーザは、初期集合から動的にクロールを進めるためのヒントとして、個々のサイトとユーザが指定したトピックとの関連性を計算するために用いられるキーワードを指定する。クローラ１０は、文書構造解析部１２により初期サイト（初期サイト集合）に対応するウェブページの文書構造解析を行い、重要度計算部１３によりクローリングの際の戦略に応じたウェブサイト（ＨＴＭＬ（Hypertext Markup Language）文書の記述上ではアンカータグ）の重要度を計算した上で、これらの情報を用いてクローリング実行部１４によりウェブサイトを検索し取得する。

【００１９】

本実施の形態の文書構造解析部１２による文書構造解析は、ウェブページのソースコードであるＨＴＭＬ文書においてブロックを識別する。

ここでブロックとは、特定の意味を持つ情報要素のまとまりであり、ウェブページを記述するＨＴＭＬにおけるブロックレベルとは必ずしも一致しない。このブロックに含まれる情報要素は、ブロックの属性（OBJECT_LIST）に、情報要素のリストとして登録され、メインメモリ１０３やＣＰＵ１０１のキャッシュメモリに保存される。このブロック化により、ＨＴＭＬ文書に含まれる情報要素のう

ち、相互に関連する情報要素が対応付けられることとなる。

情報要素には、単一要素と複数の単一要素をマージして構成された複合要素とがある。解析の最初の時点では、全ての情報要素は単一要素として識別され、解析を進めることによって複数の情報要素が複合要素としてマージされる。この情報要素は、以下の属性を持つ。

・TYPE：情報要素のタイプ。

OBJECT_ANCHOR：アンカー。

OBJECT_TEXT_BLOCK：テキスト。

OBJECT_IMAGE：メディアのタイプ。他にAUDIO、VIDEOなどが定義可能。以降の説明では、全てのメディアタイプを代表してOBJECT_IMAGEを記述する。

OBJECT_DELIMITER：情報要素のTYPEによらず区切り記号としての役割であると解析された場合に指定される。

・URL：ＵＲＬ。

OBJECT_ANCHORの場合は、HREFで指定された値。

OBJECT_IMAGE などの場合は、SRCなどで指定された値。

・TITLE：タイトル。

OBJECT_ANCHOR の場合は、Ａ（アンカー）タグで囲まれたテキスト部分。

OBJECT_IMAGE などの場合は、ALTなどで指定されたテキスト部分。

・DESCRIPTION ：記述。

OBJECT_TEXT_BLOCK の場合は、そのテキスト部分。

OBJECT_ANCHOR の場合は、関連するテキストが得られた場合、それらをマージする際にテキスト記述として追加される。

・REFERRER：参照情報。

OBJECT_ANCHOR，OBJECT_TEXT_BLOCK の場合、それに関連する OBJECT_IMAGE などの他のメディアタイプの情報要素が追加される。

・EMPHASIS：強調表現。

その情報要素が強調的な表現をされているかどうかが指定される。

【００２０】

図４は、文書構造解析部１２によるブロック化のアルゴリズムであるウェブペ

ージの文書構造解析処理の手順を説明するフローチャートである。

図４を参照すると、文書構造解析部１２は、まず解析対象のウェブページのＨＴＭＬ文書を読み込み、そのタグの記述に基づいて暫定的なブロック解析を行う（ステップ４０１）。ここでは、ＨＴＭＬ文書において、一般的に意味が区切られる箇所に記述されることの多いタグを手掛かりにして、処理対象のＨＴＭＬ文書を仮にブロックごとに区分する。このブロック解析に用いられるタグの例としては、例えば“ＢＬＯＣＫＱＵＯＴＥ”、“ＤＤ”、“ＤＩＶ”、“ＤＬ”、“ＦＯＲＭ”、“Ｈ１”、“Ｈ２”、“Ｈ３”、“Ｈ４”、“Ｈ５”、“ＨＲ”、“ＬＩ”、“Ｐ”、“ＴＡＢＬＥ”、“ＴＤ”、“ＴＨ”、“ＴＲ”、“ＵＬ”などが挙げられる。すなわち、これらのタグが出現した場合、ブロックの開始とする。そして、それぞれのタグに対応する終了タグ（＜／ＴＡＢＬＥ＞など）が出現した場合は、そのブロックの終了とする。

また、上記のようにして決定されたブロック内に、次のようなタグ、あるいは、テキストが出現した場合は、そのブロック中に上述した情報要素を作成する。

Ａ（アンカー）タグ：OBJECT_ANCHORを生成する。URLとして、HREFで指定された値。TITLEとして、Ａタグで囲まれたテキストを指定する。

タグで囲まれていないテキスト部分：OBJECT_TEXT_BLOCKを生成する。TITLEとして当該テキストを指定する。ＩＭＧなどメディアを指定するタグ：OBJECT_IMAGEなどを生成する。ＩＭＧタグの場合、URLに、SRCで指定された値を、TITLEとして、ALTで指定された値を指定する。

さらに、これら個々の情報要素を作成する際に、その情報要素が強調表現であるかどうかを判断し、そうであれば、属性EMPHASISにその強さを指定する。強調表現であるかどうかは、例えば、“Ｂ”、“ＣＥＮＴＥＲ”、“ＥＭ”、“ＳＴＲＯＮＧ”、“ＴＨ”、“Ｕ”のようなタグによってその情報要素が囲まれているかどうかによって判断する。

以上のようにして得られたＨＴＭＬ文書の構造データは、図１に示したメインメモリ１０３やＣＰＵ１０１のキャッシュメモリに一時的に格納される。

【００２１】

次に、文書構造解析部１２は、ステップ４０１で作成された構造データをメイ

ンメモリ１０３等から読み出し、この構造データに含まれる情報要素のうちで不要な情報要素を識別する（ステップ４０２）。URLが同じであるOBJECT_IMAGE、あるいは、同じTITLEを持つOBJECT_ANCHOR、あるいは、同じDESCRIPTIONを持つOBJECT_TEXT_BLOCKが複数出現されている場合は、情報を運ぶという意味合いは弱く、単なる区切り記号的な役割を担う。したがって、不要な情報要素といえる。例えば、図５（Ａ）に示すようなイメージタグは、ＬＩタグと同様の区切りの役割を果たしている。また、図５（Ｂ）に示すようなアンカータグは、情報を運ぶ意味合いは少ない。

ＨＴＭＬ文書中に出現する情報要素の数を数えながら、このような情報を運ぶ意味合いの少ない情報要素を識別し、そのタイプをOBJECT_DELIMITERとしてマークする。具体的には、不要イメージに関しては、以下のアルゴリズムによってマークを行う。なお、イメージの同一性は、URLが同じかどうかで判断する。

・１つのブロックが１つのイメージのみを含む場合は、OBJECT_DELIMITERの候補とする。

・１つのブロックが複数の同一イメージのみを含む場合は、OBJECT_DELIMITERの候補とする。

・文書全体において、複数のブロックで候補となっているものに対し、OBJECT_DELIMITERをマークする。

また、不要なOBJECT_ANCHOR、及びOBJECT_TEXT_BLOCKに関しては、同じタイトルを持つ要素が予め設定された所定の閾値以上あれば、それらをOBEJCT_DELIMITERとしてマークする。かかる閾値を用いた判断としては、例えば、テキスト長が、比較的短く（１２バイト以下）、かつ、出現回数が３回以上などであれば不要な要素とみなす。また、同じ文書内に非常に多く出現した（例えば１０回以上）全く同じOBJECT_ANCHOR，OBJECT_TEXT_BLOCKも不要な要素とみなす。これらの閾値は、システムの動作環境や用途などに応じて、経験的にあるいは適当な回数の実験を経て好ましい値を求めることができる。上記の例では、テキスト長は、実際のインターネット上のウェブページを対象として観察した結果から経験的に求めたものである（例えば、“先頭に戻る”、“戻る”、“キャッシュ”などがこの例に相当する）。また、出現回数の閾値も同様に実際に観察した結果から経験

的に求めたものである（例えば、購買サイトの“購入はこちら”などのアンカー）。

以上のようにして不要と判断されたアンカーは、クロールの対象とならない。

【００２２】

次に、文書構造解析部１２は、構造的に意味のないブロックを削除する（ステップ４０３）。例えば、所定のブロックについて、そのブロック自身が情報要素を持たず、かつ内包するブロック（子ブロック）が１つの場合は、当該ブロックを下位ブロック（すなわち当該子ブロック）とマージする。

【００２３】

次に、文書構造解析部１２は、情報要素リストのマージ、ブロック分割及びリストタイプの識別を行う（ステップ４０４）。まず、同一ブロック内の情報要素をマージする。情報要素のマージにより、関連する複数の要素が１つの複合要素とされる。以下、具体例を挙げる。

OBJECT_ANCHORは、OBJECT_ANCHOR、OBJECT_TEXT_BLOCK及びOBJECT_IMAGEとのマージが可能であり、以下に示すような操作にてマージする。

・OBJECT_ANCHORとOBJECT_ANCHORとのマージ：

２つの情報要素のURLが同一の場合のみ行う。EMPHASIS属性が付与されている方を重要とみなし、そのTITLEを採用する。そして、EMPHASIS属性が付与されていない方は、そのTITLEをDESCRIPTIONに追加する。

・OBJECT_ANCHORとOBJECT_TEXT_BLOCKとのマージ：

OBJECT_ANCHORのDESCRIPTIONにOBJECT_TEXT_BLOCKのDESCRIPTIONを追加する。

・OBJECT_ANCHORとOBJECT_IMAGEとのマージ：

OBJECT_ANCHORのREFERRERにOBJECT_IMAGEを設定する。OBJECT_ANCHORのDESCIPRTIONにOBJECT_IMAGEのTITLEを追加する。

また、OBJECT_TEXT_BLOCKは、OBJECT_ANCHOR、OBJECT_TEXT_BLOCK及びOBJECT_IMAGEとのマージが可能である。なお、OBJECT_ANCHORとのマージの場合の操作は上述したOBJECT_ANCHORどうしの場合と同様であり、OBJECT_IMAGEとのマージの場合の操作は上述したOBJECT_ANCHORとOBJECT_IMAGEとのマージの場合と同様で

ある。OBJECT_TEXT_BLOCKとOBJECT_TEXT_BLOCKとのマージの場合は、次に示すような操作にてマージする。

・OBJECT_TEXT_BLOCKとOBJECT_TEXT_BLOCKとのマージ：

一方のOBJECT_TEXT_BLOCKのDESCRIPTIONに他方のOBJECT_TEXT_BLOCKのDESCRIPTIONを追加する。

基本的には、OBJECT_ANCHOR、OBJECT_TEXT_BLOCK、OBJECT_IMAGE が同一ブロック内に存在すれば、それらをまとめられるものと判断する。OBJECT_ANCHORに関しては、同じＵＲＬを参照している要素が近くにあれば、その中に含まれる要素をマージして１つの複合要素とする。

【００２４】

また、同一ブロックに情報要素が３つ以下（OBJECT_DELIMITERを含む場合は４つ以下）の場合、どのようにマージさせるかは容易に判断できるが、複数の情報要素がリストで並んでいるような場合は、情報要素リスト内の構造解析を行い、要素のマージ、あるいは、ブロックの分割を行う。情報要素リスト内の構造解析は、原則的にはＮグラム統計を用いて次のように行う。

すなわち、１グラムから解析を行い、それぞれのグラム数において、同一ブロック内で支配的な要素の並びが見つかった場合、その並びで分割を行う。例えば、３グラム統計を行った場合、OBJECT_DELIMITER、OBJECT_ANCHOR、OBJECT_TEXT_BLOCKの当該ブロック図全体に対する割合（カバー率）が一定の閾値より大きい場合（例えば８０％以上）は、その並びを分割し、情報要素をマージする。具体的には、図６に示す例では、２つの複合要素が作成される（bullet.gif（破線で囲んだ箇所）はOBJECT_DELIMITERとして解析されている）。閾値を必ずしも１００％としないのは、ブロック内に要素が列挙して書かれている場合でも、個々の要素に関しては、イメージがあったりなかったりするなどの揺れがあり、それを吸収するために経験的に閾値を決める必要があるためである（図６の例では、OBJECT_DELIMITER、OBJECT_ANCHOR、OBJECT_TEXT_BLOCKの並びは２回出現しており、それは全体をカバーしているので、カバー率はこの場合１００％である）。

【００２５】

ただし、Ｎグラム解析だけでは、解析できない場合がある。例えば、図７に示

す例のように、参考情報（破線で囲んだ箇所）が指定されている場合である。このような場合は、特にOBJECT_ANHOR、OBJECT_TEXT_BLOCKに関する対応付けがNグラム統計では行うことができない。よって、情報要素に付与されているTITLE、DESCRIPTIONのテキスト部分からキーワードを抽出し、複数の情報要素間のキーワードリストの一致度（割合や数）を調べ、一致度が一定の閾値以上である場合は当該情報要素をマージする。例えば、双方のタイトルから抽出されたキーワードの一致する割合が、共に７０％以上であるならば、マージする。一致度の計算の際にはキーワードの分類を考慮して、人名、組織名などの重みをより重くして一致度を計算することも可能である。例えば、ニュースなどの情報であれば特に人名、組織名が一致している場合その関連度は強い場合が多い。

【００２６】

また、図８に示す例では、複数の意味的なまとまりが１つのブロック内に存在する。破線で囲まれた各部分（“ＩＢＭ関連リンク”、“日本ＩＢＭ関連リンク”）がそれぞれ意味的なまとまりを構成している。このような場合は、タイトルとなる情報要素を区切りとして情報要素リストをブロックに分割する。タイトルとなる情報要素を識別するために、EMPHASIS属性を用いる。すなわち、一定の閾値以上の長さの情報要素リスト内にEMPHASIS属性が指定されているものが複数あり、その後の情報要素が類似した並びを持っている場合は、それを複数に分割する。この際、分割されたブロックのタイトルとして、EMPHASIS属性のタイトルを指定する。このような処理によって、図８のＨＴＭＬ文書に対して、図９に示すようなブロック分割がなされる。

一般に、情報要素リストの先頭の情報要素にEMPHASIS属性が付与されており、情報要素リスト中のOBJECT_ANCHORが一定の閾値以上の割合である場合（例えば８０％以上）は、当該EMPHASIS属性のタイトルをブロックのタイトルとして指定する。あるいは、先頭のみがOBJECT_TEXT_BLOCKであり、他の要素が一定の閾値以上（例えば８０％以上）のOBJECT_ANCHORである場合も同様に、先頭のOBJECT_TEXT_BLOCKのDESCRIPTIONをブロックのタイトルとする。これらの閾値は、システムの動作環境や用途などに応じて、経験的にあるいは適当な回数の実験を経て好ましい値を求めることができる。

【００２７】

このようにして抽出されたブロックのタイトルは、個々の情報要素に対する重要度の計算に用いられる。例えば図８、９に示した例で、トピックが“日本ＩＢＭ”であった場合、そのブロックに含まれる３つのアンカー“Home”、“プロダクト＆サービス”、“サポート＆ダウンロード”は、文字列“日本ＩＢＭ”を含まないが、ブロックのタイトルに含まれることによって関連する情報要素として判断される。このように文書構造を詳細かつ正確に解析することによって、単に位置的に近傍のテキストのみからでなく、離れた位置の要素間の依存関係を利用することが可能となる。

また、抽出されたブロックの情報要素リストにおいて、OBJECT_ANCHORの割合が多い場合は、当該ブロックのリストタイプを次のように決める。

・SITE_MAP：リンク総数に対して、OBJECT_ANCHORのホスト名が異なるリンクの数の割合が所定の閾値より小さい（例えば５０％以下）場合。それ以外の場合は、ホスト名に対して固有組織名を識別するための最小の文字列を求め、その文字列の異なるリンクの割合が所定の閾値より小さい場合（例えば１０％以下）は、SITE_MAP とする。これらの閾値は、例えば、インターネット上の実際のウェブページに基づいて経験的に求められる。厳密にホスト名の一致を取るだけでは、たとえ同じ会社が提供しているページであっても異なる場合がある。ページの機能によって個別のホストを用いている場合があるからである。例えば、Yahoo! Japan(http:///www.yahoo.co.jp) は、オークションには auctions.yahoo.co.jp, 旅行関連は、travel.yahoo.co.jp などホスト名を分けている。この場合は、固有組織を識別するための最小の文字列(yahoo.co.jp)の一致度によって、閾値に基づく判断を行うことができる。

・LINK_LIST：SITE_MAPの条件を満たさない場合。

【００２８】

次に、文書構造解析部１２は、以上のようにして得られたブロックごとの情報に基づいて、ブロック構造のマージを行う（ステップ４０５）。

関連する情報要素は、必ずしもブロック内に連続して出現するわけではない。すなわち、ステップ４０１やステップ４０４の処理でブロックの設定を細かくし

過ぎてしまう場合がある。そこで例えば、所定のブロックの子ブロックにおいて並列構造があるならば、それを当該所定のブロック（親ブロック）の情報要素リストとすることにより、構造をマージする。マージする方法は、ステップ４０４で説明した情報要素リストに対するマージのアルゴリズムと同様である。

【００２９】

文書構造解析部１２は、以上説明したステップ４０３乃至ステップ４０５の動作を、処理対象のＨＴＭＬ文書に対して適用が可能である限り繰り返して適用し、解析結果である構造データの構造が変更されなくなったならば、当該構造データをメインメモリ１０３やＣＰＵ１０１のキャッシュメモリに格納して構造解析の処理を終了する（ステップ４０６）。

以上のようにして、ＨＴＭＬ文書の文書構造が意味のまとまりごとにブロック化され、ＨＴＭＬ文書から抽出された情報要素がブロックの属性として記述されることにより、相互に関連する情報要素が対応付けられて当該ブロック内のアンカーに付加されることとなる。

【００３０】

次に、重要度計算部１３によるクローリング先のウェブサイト（すなわちＨＴＭＬ文書におけるアンカータグ）の重要度の計算について説明する。

図１０、１１は、重要度計算部１３により重要度を計算し、クローリング実行部１４によりウェブサイトのコンテンツを取得する動作を説明するフローチャートである。

本実施の形態において、重要度計算部１３の重要度の計算に用いられる基本的なアルゴリズム（図１０、１１にて説明されるアルゴリズム）は、下記文献２に開示されたFish-Searchと呼ばれる手法や、上述した文献１に開示されたShark-Searchと呼ばれる手法に基づく。

文献２：P. De Bra, G.-J Houben, Y. Kornatzkt, and R. Post, Information retrieval in distributed hypertexts, in Proceedins of RIAO'94, Intelligent Multimedia, information Retrieval systems and managements, New York, NY, 1994.

ただし、本実施の形態による手法は、ユーザにより指定された戦略及び文書構造解析部１２により解析されてアンカーに付加された情報要素に基づいて検索対象であるウェブサイトのスコア（重要度）を計算し、このスコアに応じてクロールの対象を動的に決定する点、及びユーザに対してキーワードリストを提示することによって、よりユーザの目的にあったウェブサイトを収集するためのインタラクションを可能にした点が拡張されている。

また、クロールの候補となるのは、文書構造解析部１２による解析の結果として抽出された、意味のある複合情報要素のうち、他の文書への参照を持つもののみである。

【００３１】

図１０、１１を参照すると、まず、ユーザが本実施の形態による情報検索システムを構成するコンピュータ装置の入力手段を操作することにより、パラメータの設定及び初期設定が行われる（ステップ１００１）。具体的には、初期ノード集合（初期サイト、以下、クロールするウェブサイトをノードと称す）、探索幅（width）、探索の深さ（Ｄ、depth）、初期ノード集合のサイズ（Ｓ）、時間制限、探索用のキーワード（Domain Query、Focused Query）、戦略（STSET）などを設定する。戦略の設定においては複数の戦略を選択することが可能であり、各戦略に対して重み付けを行うことができる。また、クローリングの回数として０（crawlingCount = 0;）がセットされる。

次に、重要度計算部１３が、初期ノード集合の個々のノードの深さをパラメータＤにセットし、それらを空のリスト（以下、ノードリスト）に挿入する（ステップ１００２）。また、メインメモリ１０３等から文書構造解析部１２による解析結果である構造データを読み出す。そして、ノードリストが空でなく、処理されたノードの数がパラメータＳより小さく、かつ時間制限内である間、次の処理を繰り返し実行する（ステップ１００３）。

【００３２】

まず、クローリングの回数を１加算（crawlingCount += 1;）する（ステップ１００４）。ここで、crawlingCountが一定の増分を越えた場合（例えば１００

サイトごと）、かつstrategicScoreの計算が大域的である場合に、strategicScoreの再計算を行い、ノードリスト中のスコア（個々のノードのスコア、すなわち各ノードであるウェブサイトにリンクするアンカーの重要度）の値を置き換える（ステップ１００５）。また、後述するように関係するキーワードが抽出されるものに関しては、それらのキーワードリストを提示してユーザによる選択を促す。ユーザがキーワードの選択を行った場合は、選択されたキーワードに応じてトピックを更新する。この間、クロールは一旦中断しても構わないが、続けることも可能である。

次に、ノードリストから先頭のノードを取り出し、カレントノードとする。このカレントノードは、図１に示したメインメモリ１０３やＣＰＵ１０１のキャッシュメモリに保持され、クローリング実行部１４にて読み出される。そして、クローリング実行部１４がネットワークインターフェイス１０６を介してインターネットにアクセスし、当該カレントノードのＵＲＬを持つコンテンツ（ウェブページや種々のデータ等）を取得する（ステップ１００６）。取得したコンテンツは、図１に示したメインメモリ１０３やハードディスク１０５などの記憶装置に格納される。

【００３３】

次に、重要度計算部１３は、カレントノードに関する探索の深さ（パラメータD）を調べ、If　depth　＞　０ならば、以下の手順でカレントノードの関連性を計算する（ステップ１００７）。

まず、カレントノードからリンクされているノード（以下、子ノードと称す）のスコア（child_node.inherited_score）を計算する（図１１、ステップ１００８）。この計算手順は以下の通りである。

If　relevance(current_node)　＞　0

Then　child_node.inherited_score　＝　d　＊ strategicScoreForPage(STSET, current_node)

ｄは、予め定義された減衰定数。０より大きく１より小さい。

Else　child_node.inherited_score　＝　d　＊ current_node.inherited_score

次に、child_node.anchor_scoreを計算する（ステップ１００９）。

child_node.anchor_score = (relevance(anchor) + strategicScoreForAnchor(STSET, anchor))/2

そして、子ノードのpotential scoreを計算する（ステップ１０１０）。

child_node.potential_score = g * child_node.inherited_score + (1 - g) * child_node.anchor_score

ｇは、予め定義された定数。０より大きく１より小さい。

【００３４】

次に、重要度計算部１３は、カレントノードの全ての子ノードに対して以下の計算を行う（ステップ１０１１～１０１３）。

If 子ノードが優先リスト中に存在しているか。

Then

ｉ）そのノードに対するリスト中の値と今計算されたpotential_scoreの大きい方を求める。

ｉｉ）スコアを最大値で置き換える。

ｉｉｉ）子ノードをリスト中の適切な位置に移動する。

Else child_nodeにpotential_scoreを付け、リスト中の適切な位置（スコア順）に挿入する。

【００３５】

さらに、重要度計算部１３は、カレントノードの全ての子ノードに対して以下の計算を行う（ステップ１０１４）。

深さ（child_node.depth）を計算する。

If カレントノードが関連がある場合

Then child_node.depth = D

Else child_node.depth = current_node.depth - 1

If 子ノードが優先リスト中に存在している。

Then

そのノードに対するリスト中の値と今計算された深さの大きい方を求める。

その値で置き換える。

【００３６】

ステップ１０１４までの処理が終了した後、またはステップ１００７で、If depth ＞ ０でない場合は、ステップ１００３に戻り、各条件を満足する限り、クローリングの回数を１加算してステップ１００４以降の処理を繰り返す。そして、ステップ１００３のいずれかの条件を満足しない場合は、重要度計算部１３及びクローリング実行部１４による処理を終了する。

【００３７】

次に、上述したアルゴリズムにおける個々の計算方法について説明する。

・relevance(current_node)の計算方法

Domain QueryとFocused Queryは、ベクトル（トピックベクトル）で表現される。そして、これらのQuery（キーワード）とテキストの一致度は、ベクトル間の距離（内積など）で計算される。current_nodeは、そのテキスト部分をベクトルに変換し、その類似度を計算する。ユーザの指定によって、関連がないと判断されたキーワードは、トピックベクトルにおいてマイナスの重要度を持つ。これらは、以下の式で計算される。

relevance(current_node) = Similarity(current_node, Domain Query) + Similarity(current_node, Focused Query)

・relevance(anchor)の計算方法

relevance(current_node) = Similarity(TITLE, Domain Query) + Similarity(DESCRIPTION, Focused Query)

・strategicScoreForPage(STSET, current_node)の計算方法

ユーザによって指定された戦略ごとのスコアの重み付き総和によって決定される。値は０から１の間に正規化する。個々の戦略に対するスコアの計算方法（strategicScoreForPage(ST, current_node)）は後述する。

・strategicScoreForAnchor(STSET, anchor) の計算方法

ユーザによって指定された戦略ごとのスコアの重み付き総和によって決定され

る。値は０から１の間に正規化する。個々の戦略に対するスコアの計算方法（strategicScoreForPage(ST, current_node)）は後述する。

【００３８】

次に、本実施の形態によるクローリングで用いられる戦略とそのタイプ、及びstrategicScoreForPage(ST, current_node), strategicScoreForAnchor(ST, anchor)の計算方法について、例を挙げて説明する。

本実施の形態で用いられる戦略には、局所的なものと大局的なものの２つのタイプがある。局所的な戦略は、ウェブページ内の情報のみで重要度を決定できるが、大局的な戦略は、複数のウェブページを解析することによって重要度を計算する。

・ユーザが指定したトピックに近いウェブサイトを検索（局所的戦略）

この戦略は、relevance(current_node), relevance(anchor)によって計算されるものであり、図１０、１１に示した基本的なアルゴリズム内に組み込まれている。

【００３９】

・重要なウェブサイトを検索（大局的戦略）

これは、多くのウェブサイトで同じ情報が提供されていれば、その情報は重要だと見なす戦略である。同じ情報があるかどうかを複数のサイト内で調べる必要があるので大局的である。複数のウェブサイトで同じ情報が提供されているかどうかは、例えば次の文献３に開示されているような、ウェブページからヘッドライン（見出し）を抽出する公知技術を用いることによって、知ることができる。

文献３：武田、野美山、”サイト・アウトライニング-インターネットからの情報収集と可視化技術-”，情報処理，Vol. 42, No.8, 2001.

この方法では、結果として同じ事柄を言及した情報要素の集合のみを返す。例えば、図１２に示すようなヘッドラインを持つウェブサイトが検索されることとなる。なお、図１２において、Site2～4はSite 1の子ノードである。またこの方法では、抽出されたヘッドラインを構成する情報要素のテキスト部分からキーと

なるキーワードとその重みを抽出して特徴キーワードリストを生成する。図１２の例では、主要な要素となるキーワード“ロータス”、“チボリ”、“日本IBM”、“統合”とその重みとが特徴リストにリストアップされる。

ヘッドラインに含まれる新しい情報要素に対しては、対応ノードのテキスト部分から抽出されたキーワードリストと、ヘッドラインの特徴キーワードリストの距離（内積などで求められる）を計算し、その距離を重要度とする。

この方法において、strategicScoreForPage(ST, current_node)は、カレントノードのウェブページに含まれるヘッドラインの重要度の総数に基づいて計算される。そして当該ウェブページのアンカーの総数で割ることによって正規化する。

また、strategicScoreForAnchor(ST, anchor)は、現在抽出されているヘッドライン集合の全ての特徴ベクトルとの距離の最大値とする。

【００４０】

・関連するイメージなどを多く含むウェブサイトを検索（局所的戦略）

ＨＴＭＬ文書の文書構造を解析することによって、テキストと関連する他のメディアタイプ（Image、Audio、VideoあるいはＭＩＭＥ（Multipurpose Internet Mail Extensions）タイプに定義されているドキュメントファイル（例えばＰＤＦ（Portable Document Format））など）を対応付けることが可能となる。テキスト部分がトピックと関連性を持つかどうかに基づいて重要度が計算される。

この方法において、strategicScoreForPage(ST,current_node)は、カレントノードのウェブページに含まれる関連イメージの総数であり、当該ウェブページのイメージの総数で割ることによって正規化する。

また、strategicScoreForAnchor(ST,anchor)は、関連するイメージであれば、重要度を１とし、そうでない場合は重要度を０とする。

【００４１】

・重要なキーワードを含む情報に基づく検索（大局的戦略）

テキストの情報から抽出されたキーワードに基づいてクラスタリングを行い、各クラスタで重要と判断されたキーワードを多く含むかどうかによって重要度を判断する。この方法については、特開２００１－３２５２７２公報に詳細に開示

されている。

この方法において、strategicScoreForPage(ST,current_node)は、カレントノードのウェブページに含まれるホットワードを含む要素の重要度の総数に基づいて計算される。そして当該ウェブページのアンカーの総数で割ることによって正規化する。

また、strategicScoreForAnchor(ST,anchor)は、所定の要素がホットワードを含むのであれば、その重要度（０以上１以下）の値に設定し、そうでない場合は０とする。

【００４２】

・アンカーの数に基づく検索（局所的戦略）

カレントノードのウェブページ内に存在するアンカーの総数によって重要度を判断する。

この方法において、strategicScoreForPage(ST,current_node)は、カレントノードのウェブページに含まれるアンカーの総数に基づいて計算される。局所的に重要度を計算するには、例えば、リンクの数を１１段階に分け（０、１、２、３、４、５、６、７、８、９以上）、それぞれにスコア（０から１までの値）を与える。大局的に計算する場合は、検索されたウェブサイトの集合において、最大数のアンカーを含むウェブページの当該アンカーの数で正規化する。

また、strategicScoreForAnchor(ST,anchor)は、全て０とする。

【００４３】

・期間の限定に基づく検索（局所的戦略）

ウェブページ内に出現する情報の発信された期間を限定する。ＨＴＴＰプロトコルで得られる最終更新日付が、限定された期限内であれば、重要度を１とし、そうでなければ、期限からの超過日数を正規化した値を重要度とする。

・リンク集に基づく検索（局所的戦略）

ＨＴＭＬ文書の文書構造解析によって得られたリンク集に含まれるリンクの数を重要度とする。

・被参照リンクの数に基づく検索（大局的戦略）

他のウェブサイトから参照されている数を重要度とする。

・被参照リンクの参照の数（大局的戦略）

カレントノードのウェブページを参照しているリンク数を重要度とする。

【００４４】

上記のようにして、クローラ１０は、ユーザにより選択された任意の戦略を用いてウェブサイトのクロールを行い、検索されたウェブサイトの集合（以下、サイト集合）を得る。得られたサイト集合は、ステップ１００６で個々に取得されたコンテンツと共に、メインメモリ１０３やハードディスク１０５等の記憶装置に格納される。

【００４５】

ウェブサイト選別部２０は、メインメモリ１０３等の記憶装置に格納されているサイト集合の中から不要なウェブサイトを選別し、選別されたウェブサイト及びそのコンテンツを削除する。不要なウェブサイトとしては、トピックとの関連がないウェブサイト、時間条件を満たさないウェブサイトが挙げられる。

クローラ１０はユーザにて指定されたトピックを表すキーワードに基づき所定の戦略に従ってクローリングを行うのであるが、得られたサイト集合には、ユーザが指定したトピックと関連が低い、あるいは関連のないウェブサイトが含まれる可能性がある。そこで、クローリングにより得られたサイト集合に対してトピックとの一致度を求め、関連のないウェブサイトはサイト集合から削除する。ただし、クロール時に得られる参照構造で関連すると判定された複数のウェブサイトの中間に位置するウェブサイト（当該ウェブサイトを介してリンクされる場合）は、中間位置の当該ウェブサイト自体にトピックとの関連性がない場合でも削除しない。

また、サイト集合には、ユーザが指定した時間に一致しないウェブサイト（指定期限内に検索されていないウェブサイトなど）が含まれる可能性がある。そこで、クローリングにより得られたサイト集合に対してＨＴＴＰプロトコルで得られるLAST_MODIFIED属性とユーザが指定した時間条件とを比較し、一致しないウェブサイトはサイト集合から削除する。

【００４６】

レポート作成部３０は、クローラ１０にて検索され、ウェブサイト選別部２０

にて選別されたサイト集合に対して、個々の戦略に対するスコアの総得点を計算し、それらを正規化したレポートを作成し、メインメモリ１０３やハードディスク１０５等の記憶装置に格納する。例えば、イメージに関しては、１サイト当たりの関連イメージの数などが情報としてレポートに含まれることとなる。レポートは、例えばＨＴＭＬ文書として作成し、ウェブブラウザを用いて閲覧できるようにすることができる。

複数のトピックに対する戦略のスコアと比較する（平均値との比較、あるいは標準偏差を求める）ことにより、そのトピックがどのような戦略に一致しているのか（例えば、イメージが多い、情報が多い（掲示板などが多く含まれる））などの傾向を知ることができる。

【００４７】

以上のようにして、本実施の形態の情報検索システムによれば、最適な戦略を組み合わせることによって、ユーザの目的により適したサイト集合を獲得することができる。

また、クローリングするために用いられるアンカーが独立でなく、対応するテキストに対して適切に対応付けられるため、トピックとの関連性をより正確に判断することができる。

さらに、文書構造解析を行って、ＨＴＭＬ文書内で意味のないアンカーを排除するため、不要なウェブサイトをクロールする無駄を防ぐことができる。

そして、文書構造解析によって得られたブロックを利用することにより、位置的に離れた要素の依存関係をクローリングに利用できる。

このブロックを認識することによって、ウェブページ中のリンク集を特定することができるため、高品質なリンク集を収集し抽出することもできる。

この他、クロール中に（例えば図１０のステップ１００５の段階で）ユーザに対して関連キーワードを提示することによって、指定されたトピックの曖昧性を解消できる。例えば、トピック「ジャガー」は、「車」「動物」「ミック・ジャガー（ロック歌手）」などの曖昧性を持つが、クラスタリングなどによって、関連キーワードが表示し、ユーザがこの関連キーワードを指定することにより、クロール対象を絞り込むことができる。具体的には、車の「ジャガー」を検索した

い場合は、「ミック・ジャガー」、「ライブ」などの関連キーワードを指定してマイナスの重要度を与えることによって、これらの関連キーワードを含むウェブサイトをクロール対象から外すことができ、結果としてトピックの曖昧性を解消することができる。

【００４８】

図１３、１４、１５は、具体的なウェブページのＨＴＭＬ文書に対して文書構造解析を行った様子を示す図である。図１３はウェブページを表示した様子を示し、図１４は、図１３のウェブページのＨＴＭＬ文書に対して図４のステップ４０１におけるタグによるブロック解析を行った状態の構造データを示し、図１５は、ステップ４０２以降の解析処理により整理されたブロック列の状態の構造データを示す。なお、図１４、１５にはＨＴＭＬ文書の解析結果の一部のみを記載している。

図１３を参照すると、対象であるＨＴＭＬ文書は、レイアウトを揃えるためにテーブル（TABLE）タグを多用していることがわかる。そのため、これらのオブジェクト１４０１～１４０６は、表示上はまとまって見えるとしても内部構造上は離れている場合がある。例えば、画面の下部中央にある「ニュース」オブジェクトは、図１３に示すように表示上はまとまっているが、実際の構造は図１４に示すようにテーブルタグを用いて位置揃えがなされており、意味的には余分なタグが多数挿入されている。本実施の形態の文書構造解析を行うことにより、図１５に示すように、これらのオブジェクトが１つの要素１５０１として解析され、さらにそのタイトルとして「ニュース」が付与される。

図１４におけるdotted_rule_197px.gifのイメージオブジェクト１４０７は、区切りとして利用されているが、これも正しく認識されており、解析中は区切りの意味合いを持つ情報として利用されるが、図１５に示す解析後の構造には含まれない。

また、e-business hosting（ HYPERLINK "http://www.ibm.com/services/jp/webhosting/" http://www.ibm.com/services/jp/webhosting/）のオブジェクト１４０８～１４１０などのように、同じＵＲＬを指している情報要素は、本実施の形態の文書構造解析により、図１５に示すように１つの要素１５０２にまとめら

れる。

【００４９】

このように、文書構造解析部１２による解析の結果、ＨＴＭＬ文書の構造が意味のあるまとまりごとにブロック化されるため、この構造に基づいてアンカーのリンク先へ遷移し、ウェブサイトをクロールすることにより、無駄のない適切なクローリングを行うことができる。

【００５０】

【発明の効果】

以上説明したように、本発明によれば、情報の使用目的に応じて多様な戦略による柔軟な情報検索を可能とすることができる。

また、本発明によれば、ウェブページのクローリングにおいて、この多様な戦略による情報検索を実現するために、ウェブページに含まれる情報を有効に活用して検索を行うことができる。

【図面の簡単な説明】

【図１】　本実施の形態による情報検索システムを実現するのに好適なコンピュータ装置のハードウェア構成の例を模式的に示した図である。

【図２】　図１に示したコンピュータ装置にて実現される本実施の形態による情報検索システムの構成を示す図である。

【図３】　本実施の形態の情報検索システムによる情報検索の概略的な流れを示すフローチャートである。

【図４】　本実施の形態の文書構造解析部によるウェブページの文書構造解析処理の手順を説明するフローチャートである。

【図５】　本実施の形態の文書構造解析処理により解析される不要な情報要素の例を示す図である。

【図６】　本実施の形態の文書構造解析処理による情報要素のマージの例を説明する図である。

【図７】　本実施の形態の文書構造解析処理による情報要素のマージの他の例を説明する図である。

【図８】　本実施の形態の文書構造解析処理による情報要素のマージのさら

に他の例を説明する図であり、マージを行う前の状態を示す図である。

【図９】　図８の例における情報要素のマージを行った状態を説明する図である。

【図１０】　本実施の形態の重要度計算部により重要度を計算し、クローリング実行部によりウェブサイトのコンテンツを取得する動作を説明するフローチャートである。

【図１１】　本実施の形態の重要度計算部により重要度を計算し、クローリング実行部によりウェブサイトのコンテンツを取得する動作を説明するフローチャートである。

【図１２】　同じ事柄を言及したウェブサイトのサイト集合の例を示す図である。

【図１３】　ブラウザにて表示されたウェブページの例を示す図である。

【図１４】　図１３のウェブページのＨＴＭＬ文書をタグによりブロック解析した状態の構造データを示す図である。

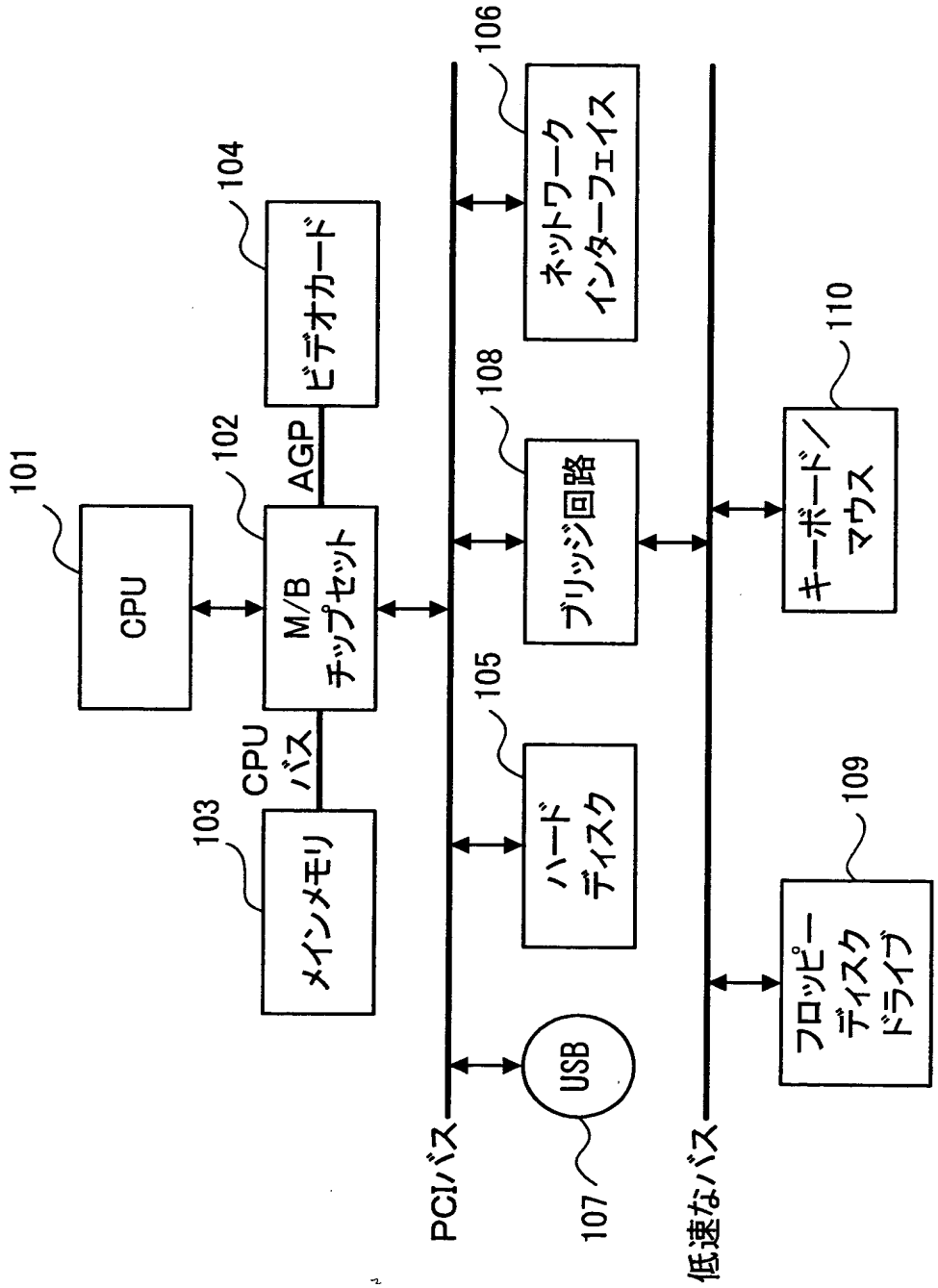
【図１５】　図１４の状態からさらに文書構造解析を行った状態の構造データを示す図である。

【符号の説明】

１０…クローラ、１１…初期サイト獲得部、１２…文書構造解析部、１３…重要度計算部、１４…クローリング実行部、２０…ウェブサイト選別部、３０…レポート作成部、１０１…ＣＰＵ、１０２…Ｍ／Ｂチップセット、１０３…メインメモリ、１０５…ハードディスク、１０６…ネットワークインターフェイス

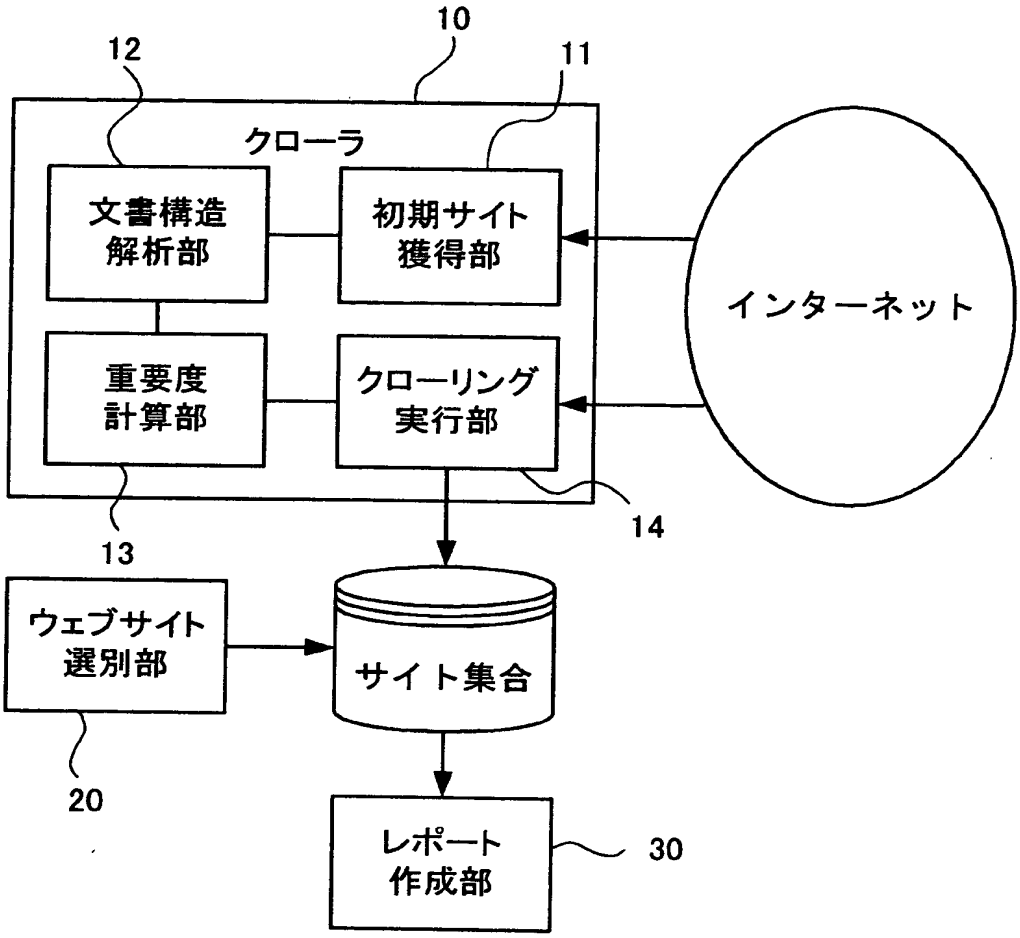
【書類名】　　図面

【図１】

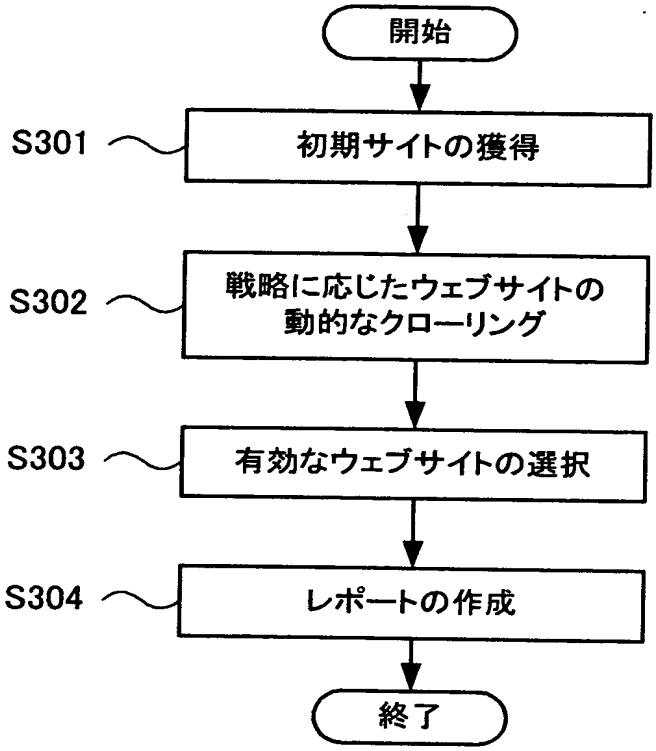


出証特２００２－３０９８２７５

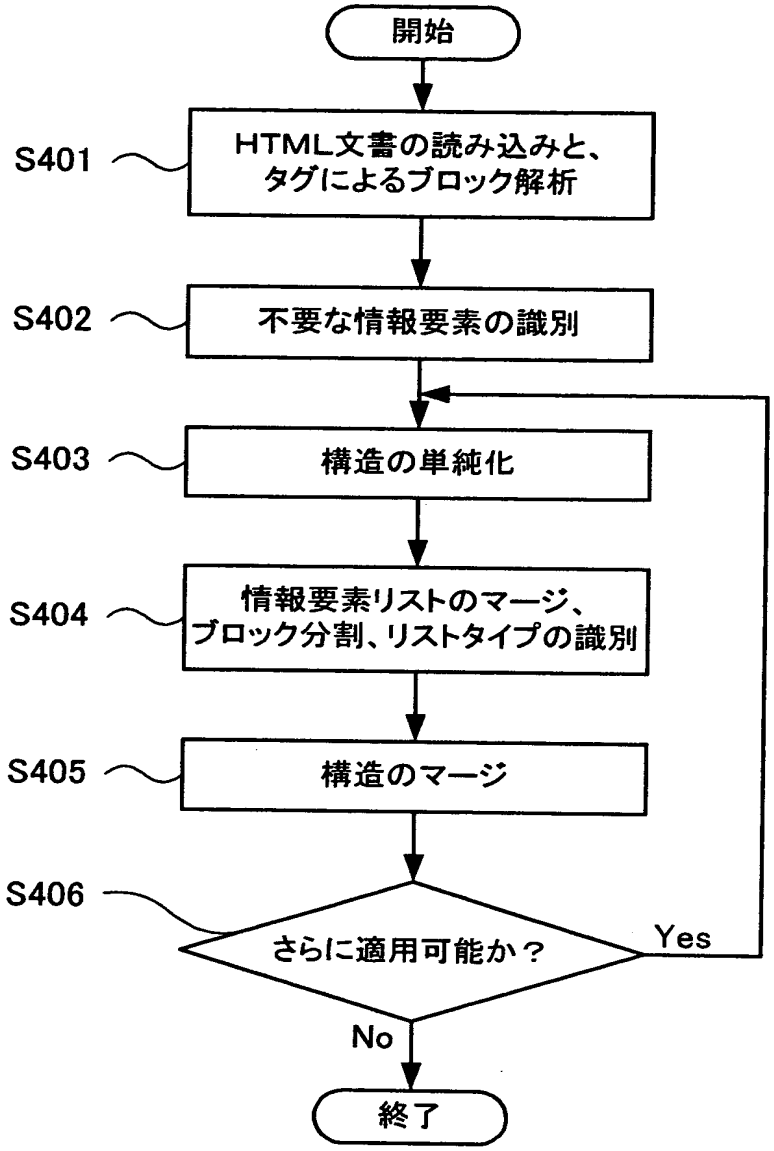
【図２】


10
12
11
クローラ
文書構造
解析部
初期サイト
獲得部
インターネット
重要度
計算部
クローリング
実行部
14
13
ウェブサイト
選別部
サイト集合
20
レポート
作成部
30

【図３】


開始
S301
初期サイトの獲得
S302
戦略に応じたウェブサイトの
動的なクローリング
S303
有効なウェブサイトの選択
S304
レポートの作成
終了

【図４】


開始
S401
HTML文書の読み込みと、
タグによるブロック解析
S402
不要な情報要素の識別
S403
構造の単純化
S404
情報要素リストのマージ、
ブロック分割、リストタイプの識別
S405
構造のマージ
S406
さらに適用可能か？
Yes
No
終了

【図５】

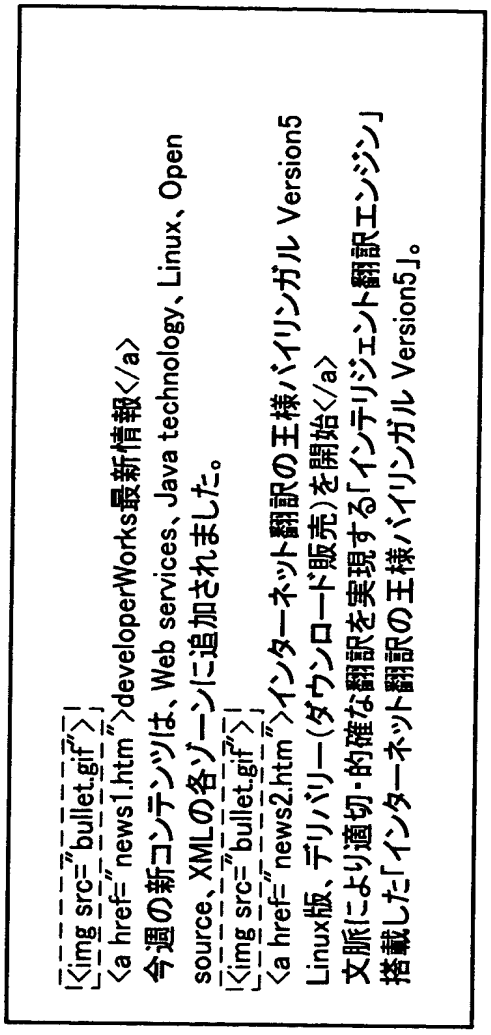
(A)

```
⋮
<img src="bullet.gif">項目1<br>
<img src="bullet.gif">項目2<br>
⋮
```

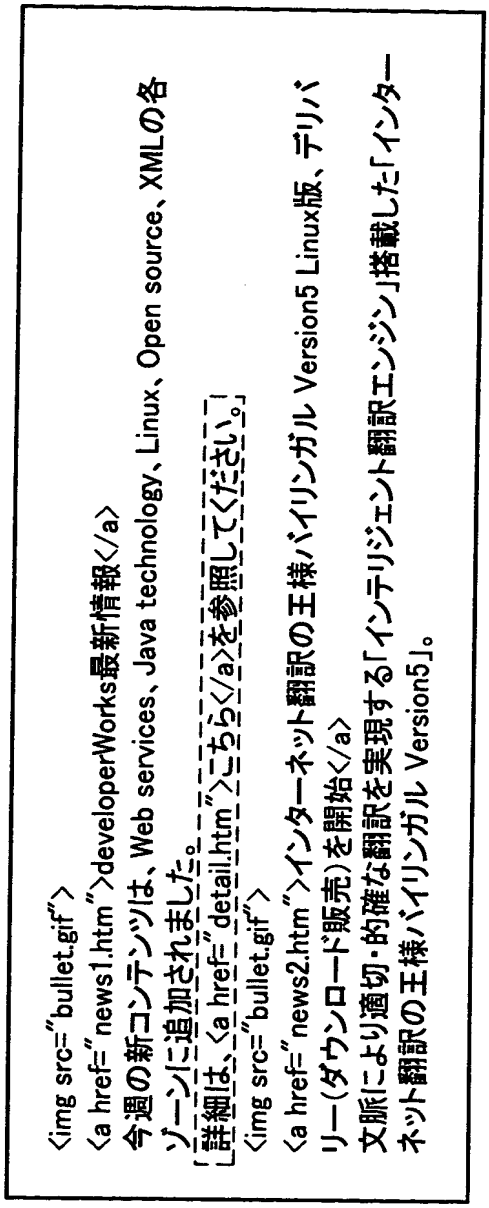
(B)

```
⋮
<a href="top.htm">先頭に戻る</a>
⋮
<a href="top.htm">先頭に戻る</a>
```

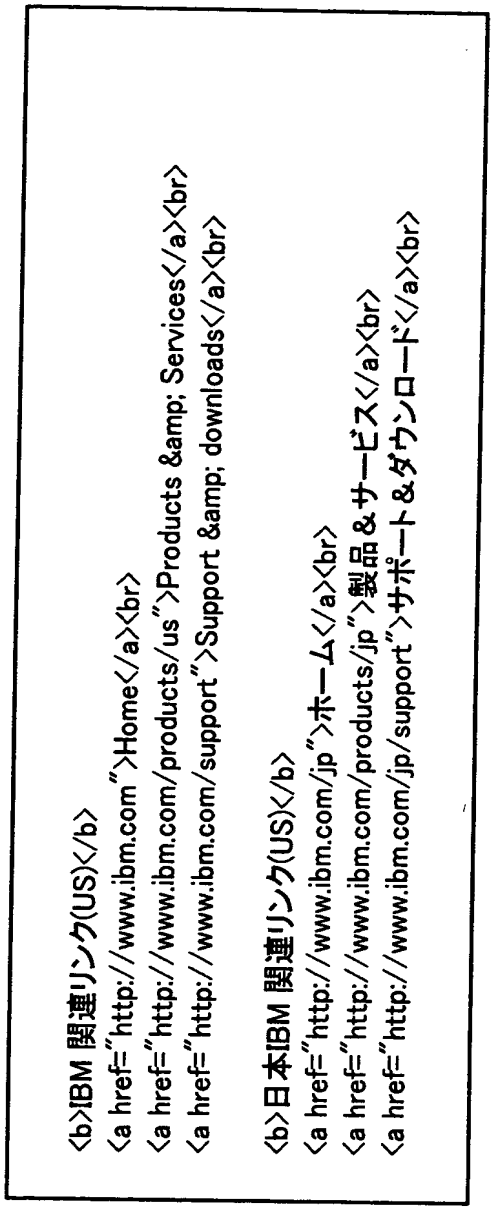
【図６】


<img src="bullet.gif">
developerWorks最新情報
今週の新コンテンツは、Web services、Java technology、Linux、Open source、XMLの各ゾーンに追加されました。
<img src="bullet.gif">
インターネット翻訳の王様バイリンガル Version5 Linux版、デリバリー（ダウンロード販売）を開始
文脈により適切・的確な翻訳を実現する「インテリジェント翻訳エンジン」搭載した「インターネット翻訳の王様バイリンガル Version5」。

【図７】


<img src="bullet.gif">
developerWorks最新情報
今週の新コンテンツは、Web services、Java technology、Linux、Open source、XMLの各ゾーンに追加されました。
「詳細は、こちらを参照してください。」
<img src="bullet.gif">
インターネット翻訳の王様バイリンガル Version5 Linux版、デリバリー（ダウンロード販売）を開始
文脈により適切・的確な翻訳を実現する「インテリジェント翻訳エンジン」搭載した「インターネット翻訳の王様バイリンガル Version5」。

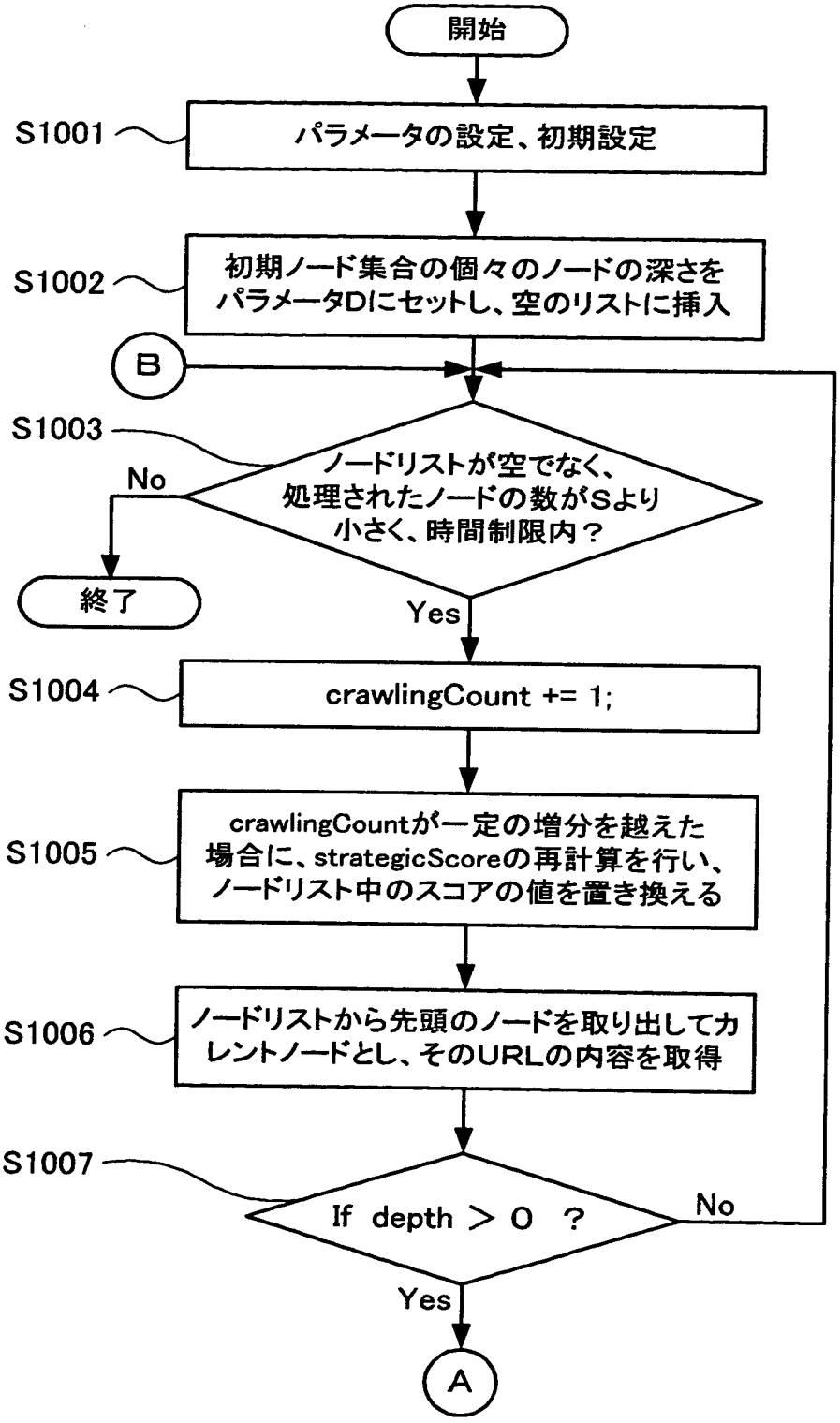
【図８】


IBM 関連リンク(US)
Home
Products & Services
Support & downloads
日本IBM 関連リンク(US)
ホーム
製品＆サービス
サポート＆ダウンロード

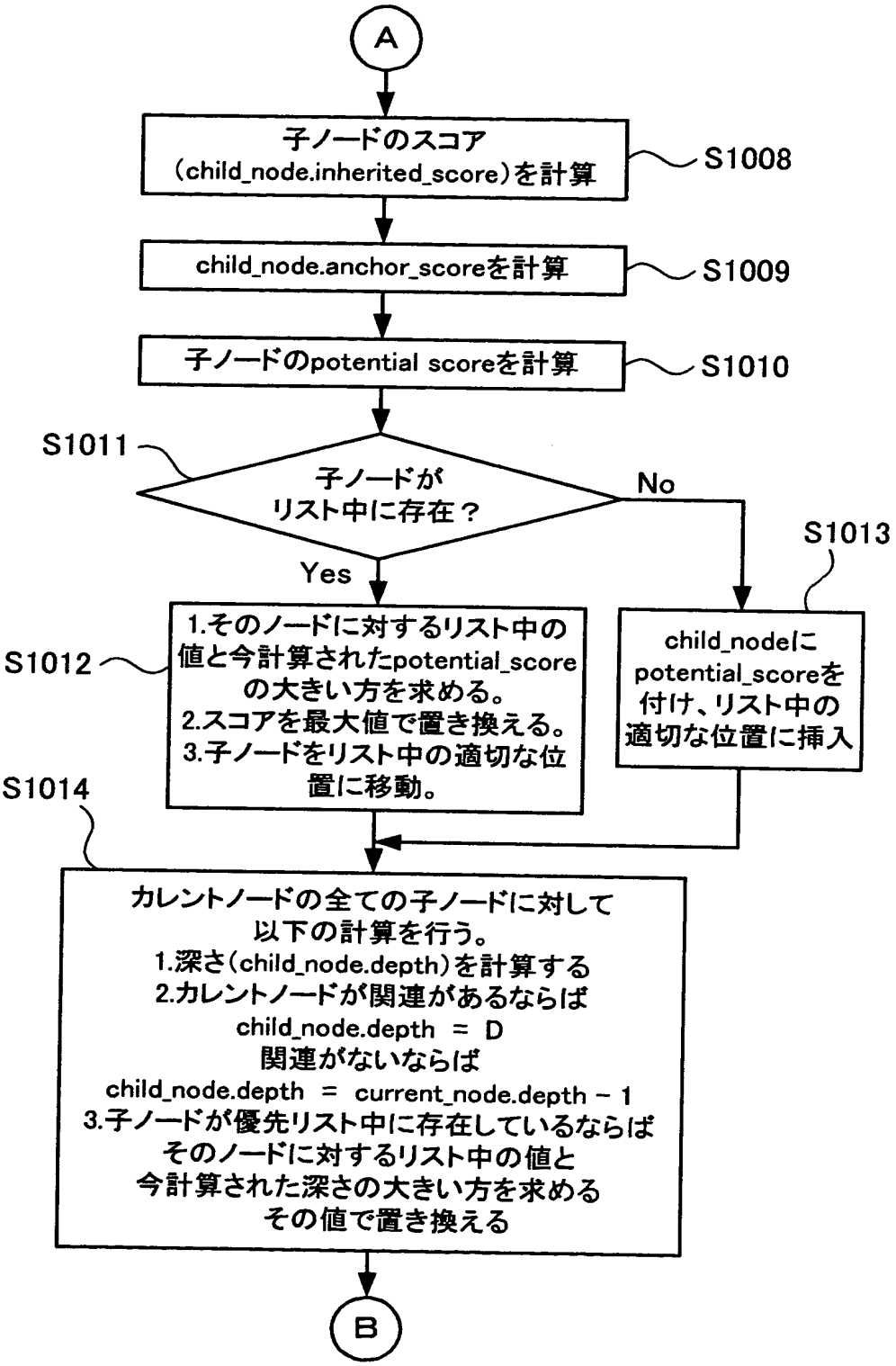
【図９】

```
<block>
<block title="IBM関連リンク">
 <object type="anchor" url="http://www.ibm.com" title="Home"/>
 <object type="anchor" url="http://www.ibm.com/products/us" title="Products &amp;amp; Services"/>

 <object type="anchor" url="http://www.ibm.com/support" title="Support &amp;amp; downloads"/>
 </block>
<block title="日本IBM関連リンク">
 <object type="anchor" url="http://www.ibm.com/jp" title="Home"/>
 <object type="anchor" url="http://www.ibm.com/products/jp" title="プロダクト＆サービス"/>

 <object type="anchor" url="http://www.ibm.com/jp/support" title="サポート＆ダウンロード"/>
 </block>
 </block>
```

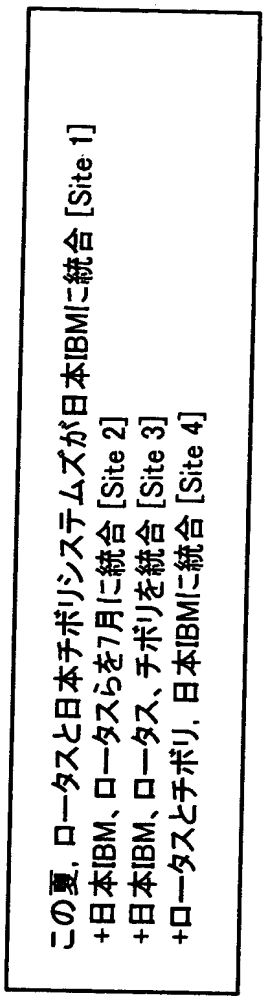
【図１０】


開始
S1001
パラメータの設定、初期設定
S1002
初期ノード集合の個々のノードの深さを
パラメータDにセットし、空のリストに挿入
B
S1003
ノードリストが空でなく、
処理されたノードの数がSより
小さく、時間制限内？
No
終了
Yes
S1004
crawlingCount += 1;
S1005
crawlingCountが一定の増分を越えた
場合に、strategicScoreの再計算を行い、
ノードリスト中のスコアの値を置き換える
S1006
ノードリストから先頭のノードを取り出してカ
レントノードとし、そのURLの内容を取得
S1007
If depth > 0 ？
No
Yes
A

【図１１】

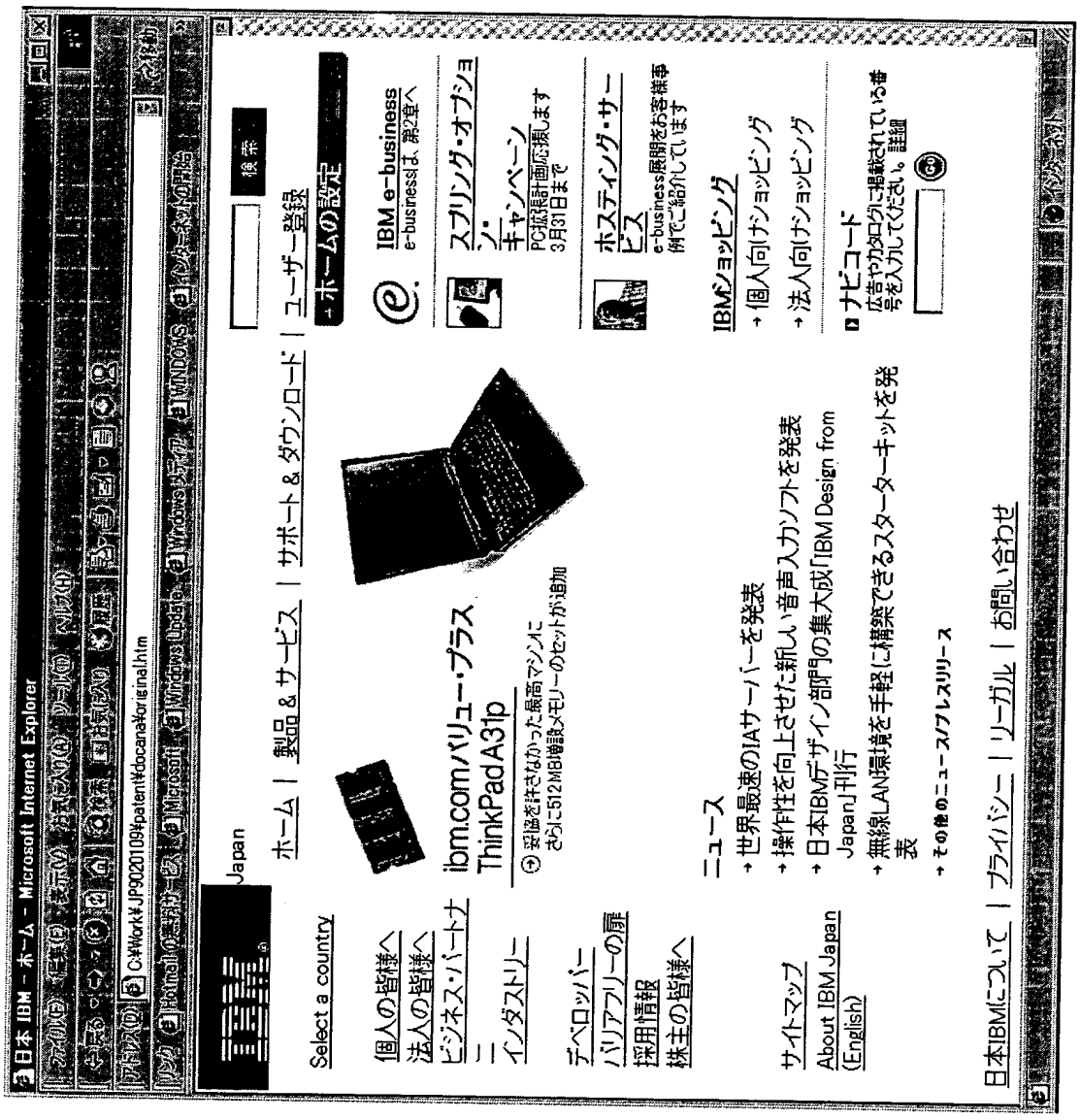

A
子ノードのスコア
(child_node.inherited_score)を計算
S1008
child_node.anchor_scoreを計算
S1009
子ノードのpotential scoreを計算
S1010
S1011
子ノードが
リスト中に存在？
No
Yes
S1013
1.そのノードに対するリスト中の
値と今計算されたpotential_score
の大きい方を求める。
2.スコアを最大値で置き換える。
3.子ノードをリスト中の適切な位
置に移動。
S1012
child_nodeに
potential_scoreを
付け、リスト中の
適切な位置に挿入
S1014
カレントノードの全ての子ノードに対して
以下の計算を行う。
1.深さ(child_node.depth)を計算する
2.カレントノードが関連があるならば
child_node.depth = D
関連がないならば
child_node.depth = current_node.depth - 1
3.子ノードが優先リスト中に存在しているならば
そのノードに対するリスト中の値と
今計算された深さの大きい方を求める
その値で置き換える
B

【図１２】

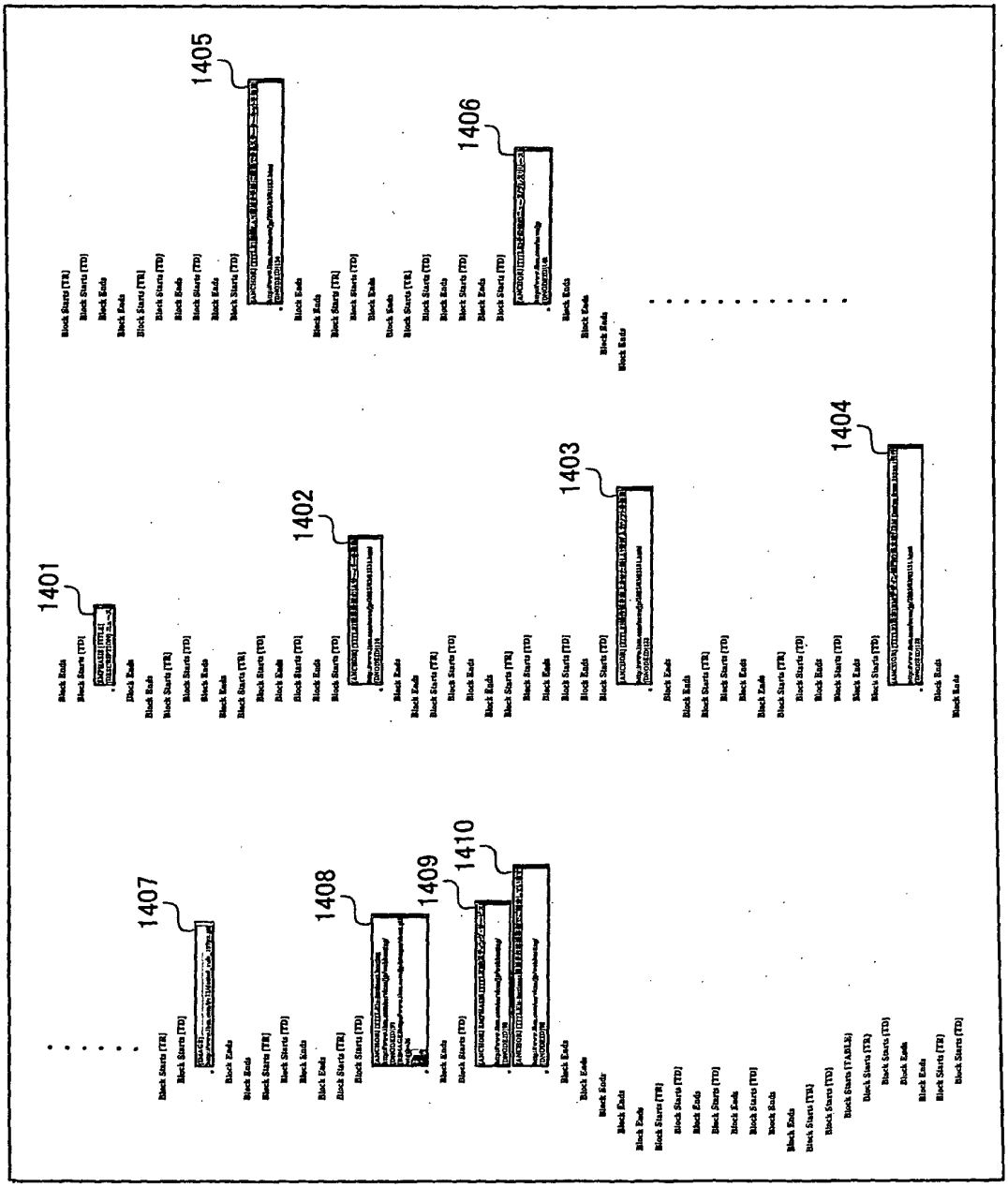


出証特２００２－３０９８２７５

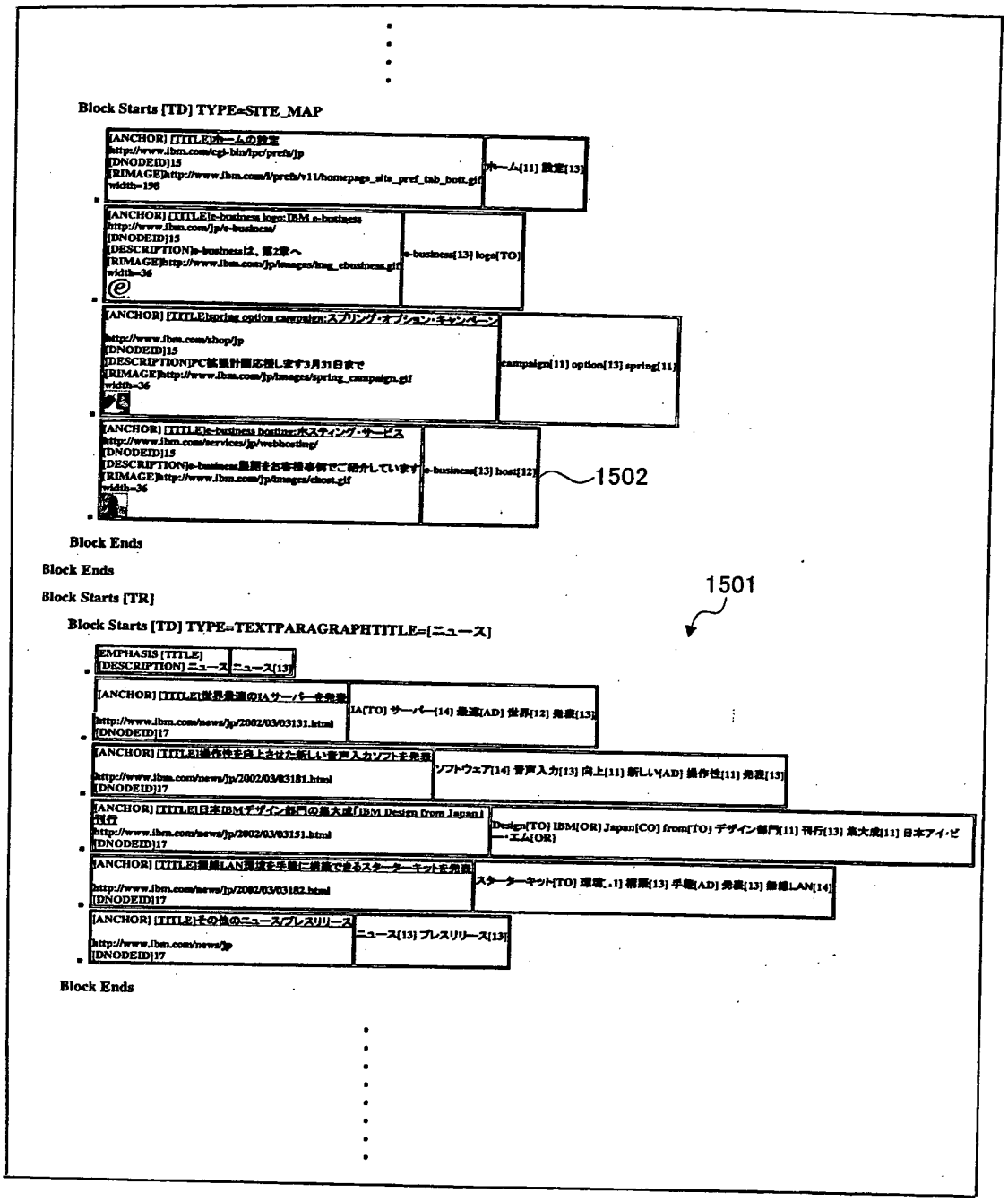
【図１３】


日本 IBM - ホーム - Microsoft Internet Explorer
C:¥Work¥JP9020109¥patent¥docana¥original.htm
Japan
ホーム | 製品 & サービス | サポート & ダウンロード | ユーザー登録
ホームの設定
検索
Select a country
個人の皆様へ
法人の皆様へ
ビジネス・パートナー
インダストリー
デベロッパー
バリアフリーの扉
採用情報
株主の皆様へ
ibm.com パリュー・プラス
ThinkPad A31p
IBM e-business
スプリング・オプション・キャンペーン
ホスティング・サービス
IBMショッピング
個人向けショッピング
法人向けショッピング
ナビコード
ニュース
世界最速のIAサーバーを発表
操作性を向上させた新しい音声入力ソフトを発表
日本IBMデザイン部門の集大成「IBM Design from Japan」刊行
無線LAN環境を手軽に構築できるスターターキットを発表
サイトマップ
About IBM Japan (English)
日本IBMについて | プライバシー | リーガル | お問い合わせ

【図１４】



【図１５】


Block Starts [TD] TYPE=SITE_MAP
[ANCHOR] [TITLE]ホームの設定
http://www.ibm.com/cgi-bin/fpc/prefs/jp
[DNODEID]15
[RIMAGE]http://www.ibm.com/i/prefs/v11/homepage_site_pref_tab_bott.gif
width=198
ホーム[11] 設定[13]
[ANCHOR] [TITLE]e-business logo:IBM e-business
http://www.ibm.com/jp/e-business/
[DNODEID]15
[DESCRIPTION]e-businessは、第2章へ
[RIMAGE]http://www.ibm.com/jp/images/tag_ebusiness.gif
width=36
e-business[13] logo[TO]
[ANCHOR] [TITLE]spring option campaign:スプリング・オプション・キャンペーン
http://www.ibm.com/shop/jp
[DNODEID]15
[DESCRIPTION]PC拡張計画応援します3月31日まで
[RIMAGE]http://www.ibm.com/jp/images/spring_campaign.gif
width=36
campaign[11] option[13] spring[11]
[ANCHOR] [TITLE]e-business hosting:ホスティング・サービス
http://www.ibm.com/services/jp/webhosting/
[DNODEID]15
[DESCRIPTION]e-business基盤をお客様事例でご紹介しています
[RIMAGE]http://www.ibm.com/jp/images/ehost.gif
width=36
e-business[13] host[12]
1502
Block Ends
Block Ends
1501
Block Starts [TR]
Block Starts [TD] TYPE=TEXTPARAGRAPHTITLE=[ニュース]
EMPHASIS [TITLE]
[DESCRIPTION] ニュース
ニュース[13]
[ANCHOR] [TITLE]世界最速のIAサーバーを発表
http://www.ibm.com/news/jp/2002/03/03131.html
[DNODEID]17
IA[TO] サーバー[14] 最速[AD] 世界[12] 発表[13]
[ANCHOR] [TITLE]操作性を向上させた新しい音声入力ソフトを発表
http://www.ibm.com/news/jp/2002/03/03181.html
[DNODEID]17
ソフトウェア[14] 音声入力[13] 向上[11] 新しい[AD] 操作性[11] 発表[13]
[ANCHOR] [TITLE]日本IBMデザイン部門の集大成「IBM Design from Japan」刊行
http://www.ibm.com/news/jp/2002/03/03151.html
[DNODEID]17
[Design[TO] IBM[OR] Japan[CO] from[TO] デザイン部門[11] 刊行[13] 集大成[11] 日本アイ・ビー・エム[OR]
[ANCHOR] [TITLE]無線LAN環境を手軽に構築できるスターターキットを発表
http://www.ibm.com/news/jp/2002/03/03182.html
[DNODEID]17
スターターキット[TO] 環境[.1] 構築[13] 手軽[AD] 発表[13] 無線LAN[14]
[ANCHOR] [TITLE]その他のニュース/プレスリリース
http://www.ibm.com/news/jp
[DNODEID]17
ニュース[13] プレスリリース[13]
Block Ends

【書類名】　　要約書

【要約】

【課題】　コンピュータを用いた情報検索において、情報の使用目的に応じて多様な戦略による柔軟な情報検索を効果的に実現する。

【解決手段】　所定のウェブページにおける意味を考慮してＨＴＭＬ文書の構造を解析する文書構造解析部１２と、この解析結果に基づき、予め定められた戦略にしたがって、このウェブページからリンクされる他のウェブサイトの重要度を計算する重要度計算部１３と、この重要度計算部１３により計算された重要度に応じてウェブサイトをクロールするクローリング実行部１４とを備える。

【選択図】　　図２

認定・付加情報

特許出願の番号	特願２００２－２１１６３４
受付番号	５０２０１０６６７４９
書類名	特許願
担当官	土井　恵子　　４２６４
作成日	平成１４年１０月　８日

＜認定情報・付加情報＞

【特許出願人】

【識別番号】　390009531

【住所又は居所】　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】　100086243

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　市位　嘉宏

【代理人】

【識別番号】　100108501

【住所又は居所】　神奈川県大和市下鶴間１６２３番１４　日本アイ・ビー・エム株式会社　知的所有権

【氏名又は名称】　上野　剛史

【復代理人】　申請人

【識別番号】　100104880

【住所又は居所】　東京都港区赤坂５－４－１１　山口建設第２ビル６Ｆ　セリオ国際特許事務所

【氏名又は名称】　古部　次郎

次頁無

出　願　人　履　歴　情　報

識別番号　　〔３９０００９５３１〕

１．変更年月日　　２００２年　６月　３日
［変更理由］　住所変更
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレーション